Biology-Oriented Synthesis

Stefan Wetzel, Robin S. Bon, Kamal Kumar, and Herbert Waldmann*

BIOLOGY – ORIENTED SYNTHESIS
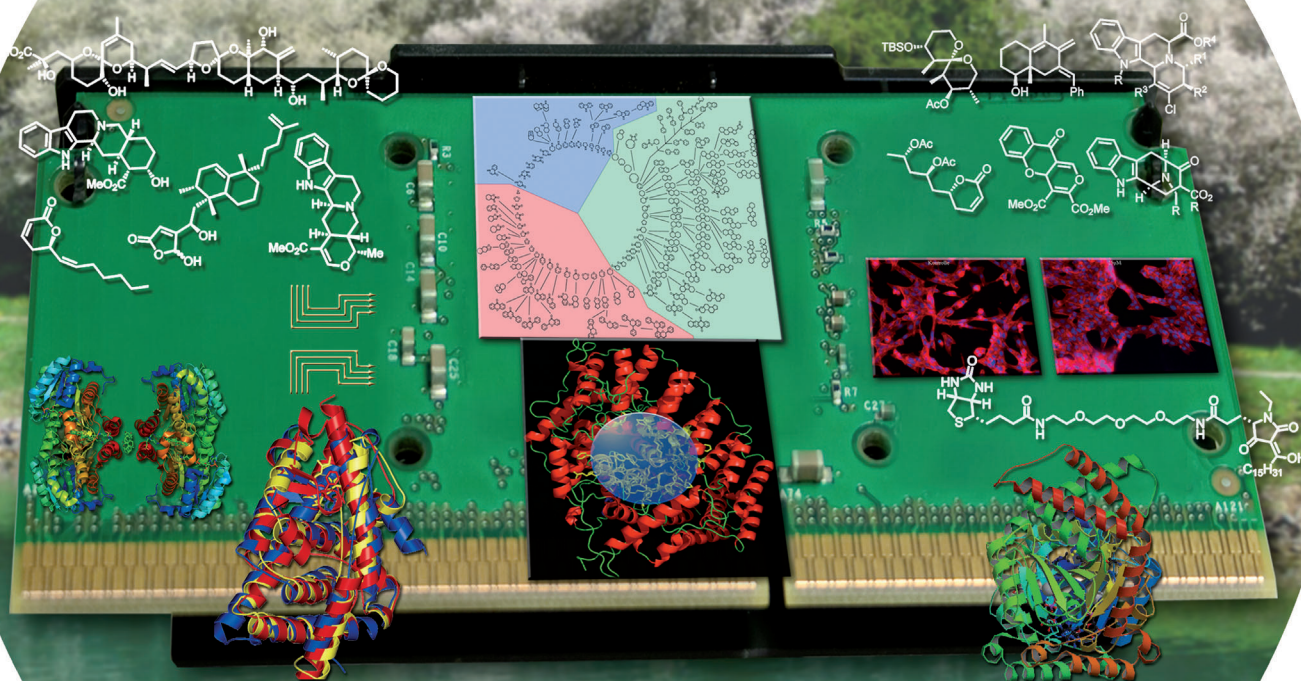
Drug Candidates
Chemical Probes
New Targets

Angewandte Chemie

*W*hich compound classes are best suited as probes and tools for chemical biology research and as inspiration for medicinal chemistry programs? Chemical space is enormously large and cannot be exploited conclusively by means of synthesis efforts. Methods are required that allow one to identify and map the biologically relevant subspaces of vast chemical space, and serve as hypothesis-generating tools for inspiring synthesis programs. Biology-oriented synthesis builds on structural conservatism in the evolution of proteins and natural products. It employs a hierarchical classification of bioactive compounds according to structural relationships and type of bioactivity, and selects the scaffolds of bioactive molecule classes as starting points for the synthesis of compound collections with focused diversity. Navigation in chemical space is facilitated by Scaffold Hunter, an intuitively accessible and highly interactive software. Small molecules synthesized according to BIOS are enriched in bioactivity. They facilitate the analysis of complex biological phenomena by means of acute perturbation and may serve as novel starting points to inspire drug discovery programs.

From the Contents

“
Nature creates nothing without a purpose”
Aristotle

1. Introduction

The interrogation of biological systems by using small molecules (substances with a molecular weight below $800 \, \mathrm{g \, mol^{-1}}$) is at the heart of chemical biology. Bioactive small molecules are excellent tools and probes for the analysis of complex biological networks and systems endowed with robust and redundant functionality. In contrast to genetic approaches, their effect is acute but not chronic. They work rapidly and reversibly, and their use is conditional and tunable (by varying their concentration).[1] Although the properties of such chemical probes often differ from those of drugs,[2] successful chemical probes are valuable sources of inspiration for drug discovery. Over the last few decades, numerous small molecules have been identified that modify the activity of a wide range of proteins, and there is a growing demand for high-quality chemical probes with clearly defined structure, potency, selectivity, mechanism of action, and availability.[3] The development of selective small-molecule modulators of all proteins encoded by the human genome has been suggested as a grand target of chemical biology research.[4] One of the main challenges in this endeavor is the identification of suitable compound classes for the perturbation of one particular protein function. Since current estimates of the number of small molecules populating druglike chemical space exceed 10^{60}, it will be impossible to investigate all the possibilities.[5] In fact, there is neither enough matter in the universe nor enough time to make them all. The key question in the development of small molecules for chemical biology research, and by analogy and extension also for drug

discovery, therefore, is how to identify the areas in chemical space that are enriched with biologically relevant compounds, that is, how to identify, map, and navigate biologically relevant chemical space?[5c]

By analogy to these limitations set for accessibility to biologically relevant small molecules, nature has been conservative in the evolution of chemical space in protein binding sites. For proteins with an average size of 300 amino acid residues and made from 20 different amino acids, more than 10^{390} unique combinations are possible.[5a] However, even the genomes of the most complicated organisms encode for only 10^4–10^5 proteins, often containing subdomains that are highly conserved within protein families. This conservatism leads to only a limited number of possible small-molecule binding sites which inspire rational approaches to ligand and inhibitor development. Current state-of-the-art methods are based, for example, on mechanistic considerations (mechanism-based inhibitors), evolutionary arguments (sequence homology), 3D protein structure (structure-based design),[6] or classification of small molecules according to predefined properties (chemical descriptors).[7] In light of the limitations set in evolution for both small molecules and protein binding sites, we have developed a conceptually alternative, structure-based approach to analyze biologically relevant chemical space and its use in the development of small molecules for chemical biology and medicinal chemistry research. We refer to this approach as biology-oriented synthesis (BIOS). BIOS is based on structural analysis of the protein and the small-

[*] Dr. S. Wetzel, Dr. R. S. Bon, Dr. K. Kumar, Prof. Dr. H. Waldmann
Max-Planck-Institut für Molekulare Physiologie
Abt. Chemische Biologie
Otto-Hahn-Strasse 11, 44227 Dortmund (Germany)
and
Technische Universität Dortmund, Fakultät Chemie
Lehrbereich Chemische Biologie
Otto-Hahn-Strasse 6, 44227 Dortmund (Germany)
E-mail: herbert.waldmann@mpi-dortmund.mpg.de

molecule world as well as the combination of structural conservatism and diversity in nature. In this Review we first delineate the philosophy behind our reasoning. We then describe the development of cheminformatic and bioinformatic methods as well as tools to identify, analyze, chart, and navigate biologically relevant chemical space, followed by the application of these methods to the design and synthesis of compound collections. Finally, we show how chemical probes can be developed according to the logic of BIOS and how they have been used to gain novel insight into biological phenomena.

2. Structural Conservatism and Diversity in Natural Product Space and Protein Binding Site Space

Small-molecule secondary metabolites created by nature (natural products) define a particularly important area of biologically relevant chemical space for bioactive small-molecule discovery. Natural products have been and continue to be a major source of inspiration for drug discovery,[8] and natural product derived and inspired compound libraries have demonstrated increased hit rates in biochemical and biological screens (for the definitions of natural product derived and inspired compounds, see Figure 1).[9] Two important properties that distinguish natural products from compounds in typical combinatorial chemistry libraries designed preferentially on the basis of chemical accessibility are their increased molecular complexity and the prevalence of stereogenic

centers. It has been shown that these molecular properties correlate with success rates as compounds transition from discovery to drugs in the clinic.[10] Natural products have evolved to interact with multiple proteins. On the one hand, the biosynthesis of natural products typically proceeds through sequential binding of biosynthetic intermediates to different enzymes. On the other hand, many natural products display a variety of biological activities, either within one organism or across species. Taken together, these arguments demonstrate that natural products most likely bind to and modulate the activities of multiple protein targets. This is particularly true for closely related analogues that define a whole class of natural products. In fact, the molecular scaffolds of natural products are highly conserved in nature (see also Section 2.1), and many natural products that share a common scaffold, but have diverse substituent patterns, display different bioactivity profiles. Therefore, the scaffolds of natural products define evolutionary-chosen "privileged structures".[11] These confer to the whole compound class the ability to interact with and bind to multiple protein targets, and, therefore, encode structural properties required for binding. As a consequence of these properties, natural product scaffolds also define biologically relevant areas of vast chemical structure space identified in evolution and nature's solution to the problem of charting and navigating it. It should be noted, however, that this solution is not exclusive, as demonstrated by the development of drugs by the pharmaceutical industry for more than a century that are not based on natural products.

Herbert Waldmann was born in 1957 in Neuwied. He received his PhD in organic chemistry in 1985 from the University of Mainz under the guidance of Prof. H. Kunz, after which he carried out postdoctoral research with Prof. G. Whitesides at Harvard University. He was appointed Professor of Organic Chemistry at the University of Bonn (1991), full Professor of Organic Chemistry at the University of Karlsruhe (1993), and Director at the MPI of Molecular Physiology Dortmund and Professor of Organic Chemistry at the University of Dortmund (1999). His research interests lie in chemical biology research, the use of small-molecule and protein probes, and microarray technology.

Stefan Wetzel completed his chemistry studies at the universities of Regensburg and Heidelberg and then joined the department of Prof. Waldmann at the Max-Planck Institute of Molecular Physiology. In his doctoral work he developed novel computational approaches for the design of focused biologically relevant libraries by using methods from the fields of cheminformatics, bioinformatics, computational chemistry as well as biochemical assays. Currently, he is a postdoctoral researcher at Novartis, working in the field of quantitative biology and computational systems biology.

Kamal Kumar obtained his PhD from G.N.D. University, Amritsar, India, under the supervision of Prof. M. P. S. Ishar. After postdoctoral research as an Alexander von Humboldt Fellow with Prof. M. Beller at Rostock, Germany in 2002, he joined the group of Prof. H. Waldmann in the Department of Chemical Biology at the Max Planck Institute of Molecular Physiology. Since May 2006 he has been leading a group in the same department. His research interests include the development of new synthetic methods towards natural product based libraries, cascade reactions, complexity-generating annulations, and probing biological functions with small molecules.

Robin S. Bon completed his PhD in organic chemistry at the Vrije Universiteit Amsterdam in 2007 with Prof. Romano Orru and carried out postdoctoral research, supported by an Alexander von Humboldt fellowship, with Prof. Herbert Waldmann at the MPI of Molecular Physiology, Dortmund. Since November 2009, he has been a senior research fellow at the University of Leeds. His research focuses on the development of small molecule modulators of protein function and tools for biochemical assays and in vivo imaging.
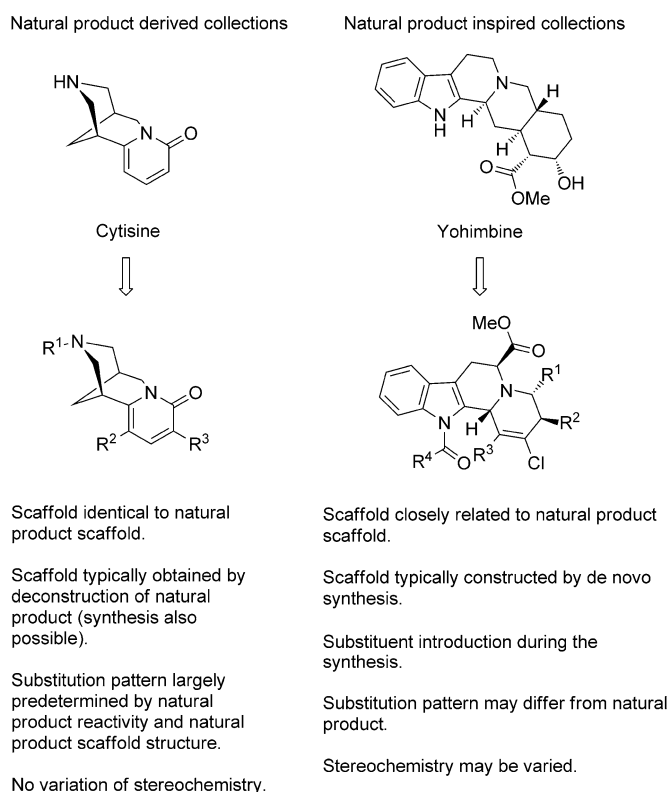
Natural product derived collections

Cytisine

Natural product inspired collections

Yohimbine

Scaffold identical to natural product scaffold.

Scaffold typically obtained by deconstruction of natural product (synthesis also possible).

Substitution pattern largely predetermined by natural product reactivity and natural product scaffold structure.

No variation of stereochemistry.

Scaffold closely related to natural product scaffold.

Scaffold typically constructed by de novo synthesis.

Substituent introduction during the synthesis.

Substitution pattern may differ from natural product.

Stereochemistry may be varied.

Figure 1. Natural product derived and natural product inspired compound libraries.

The highly selective recognition of natural products and their precursors by the biosynthetic machinery as well as specific receptors and targets requires tight molecular interactions between the natural products and ligand-binding sites of proteins. Therefore, the protein structure has to match the structural features of the natural products. The 3D structures of proteins are determined by the arrangement of secondary structural elements, such as α helices and β sheets, in the protein backbone, thereby resulting in characteristic fold types of individual protein domains joined to form the whole protein. Subfolds within protein domains also determine the sizes and shapes of their ligand-binding sites as well as the spatial arrangement of catalytic and ligand-recognizing residues. The identity and chemical nature of the amino acid residues, in particular their side chains, determines the kind of ligand that can be bound. The structure of the protein fold is conserved in nature on a higher level than the amino acid sequence, and protein domains with low sequence homology can make very similar folds. The estimated total number of fold types in nature is in the range 1000–8000 and even lower if restricted to the structures of major protein families.[12]

The recognition that nature is conservative in the evolution of both the scaffolds of natural products and protein backbones, complemented by the diversity of amino acid side chain residues in proteins and natural product substituents, has led us to propose and investigate a possible analogy between the scaffolds of the natural products and the subfolds of ligand binding sites with incorporated hotspots for binding. We hypothesized that highly conserved natural

product scaffolds match highly conserved subfolds of ligand binding sites, and that the interaction of diverse natural product substituents with diverse amino acid residues in ligand binding sites establishes selective and potent binding. In this scenario, the natural product scaffolds determine the spatial positioning of their substituents and, therefore, they fit into ligand binding sites with complementary shapes and sizes, that is, with complementary subfolds (Figure 2). However, binding will only occur if the properties and sizes of the natural product substituents and amino acid residues in the ligand binding sites match as well. According to this proposal, natural products (and possibly other small-molecules classes) with similar scaffolds are likely to bind to proteins with similar ligand binding site subfolds. Therefore, the identification of structural analogies between natural product scaffolds and protein subfolds could guide the development of natural product inspired compound libraries. Ideally, such compound libraries, based on particular natural product scaffolds and equipped with sufficient substituent diversity, would contain ligands for multiple protein domains with similar subfolds.

This reasoning puts the structure of the small-molecule ligands and the ligand-sensing protein cores into the limelight of compound discovery. It inherently reflects a chemocentric approach to compound development and conceptually is an alternative to other valid approaches based, for example, on mechanistic and evolutionary considerations or approaches aimed at maximizing chemical diversity (see above).

It is important to note that the diversity in the ligand binding sites as a result of amino acid variation at a given subfold architecture necessitates the development of natural product inspired compound collections. Only if the diversity of the substituents attached to a given natural product scaffold matches the diversity of the amino acid side chains possibly occurring in otherwise structurally similar domain subfolds, will such compound collections yield ligands for multiple proteins.

To investigate this proposal and the possibility of identifying biologically prevalidated starting points in chemical space for the generation of small-molecule libraries, chem-informatic and bioinformatic methods were developed to identify, chart, analyze, and navigate biologically relevant chemical space as well as the protein binding site space. These methods were then employed to guide the development of compound collections and to prospectively assign bioactivity for selected compound classes.

2.1. Structural Classification of Natural Products (SCONP)

Early investigations into the properties of natural products[13] were geared toward understanding the differences between natural products and typical compounds used in medicinal chemistry, and to decode the molecular parameters determining the biological relevance of natural products. Subsequently, the concept of scaffold trees was introduced and applied to the Dictionary of Natural Products (DNP). This effort resulted in the first Structural Classification of Natural Products (SCONP)[14], which effectively charted the chemical space of natural products as contained in the DNP,
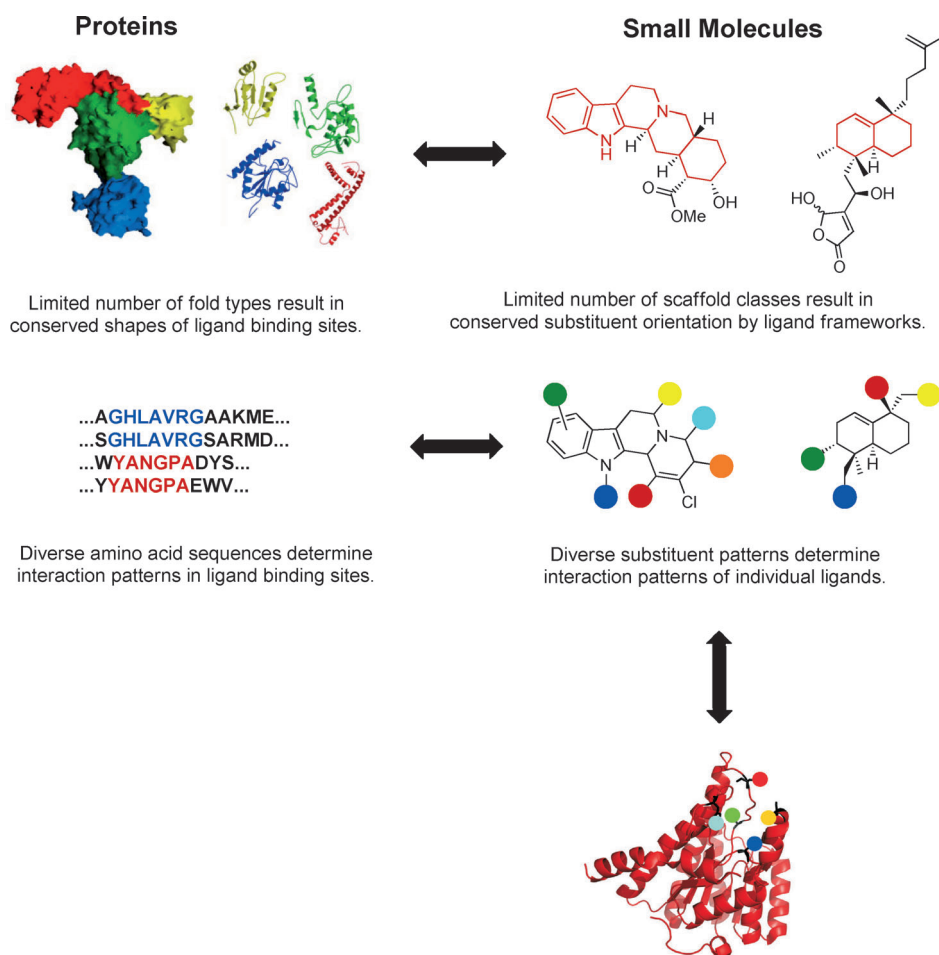
Proteins

Small Molecules

Limited number of fold types result in conserved shapes of ligand binding sites.

Limited number of scaffold classes result in conserved substituent orientation by ligand frameworks.

...A**GHLAVRG**AAKME...
...S**GHLAVRG**SARMD...
...W**YANGPA**DYS...
...Y**YANGPA**EWV...

Diverse amino acid sequences determine interaction patterns in ligand binding sites.

Diverse substituent patterns determine interaction patterns of individual ligands.

Figure 2. Scaffold-substituent analogy between small molecules and proteins. The small-molecule scaffold determines the spatial orientation of the substituents, whereas the protein subfold arranges the amino acid side chains spatially. Binding occurs when compatible substituents match in their spatial positioning so they can interact.

the most comprehensive database resource of natural product structures (Figure 3).[15]

To reduce the high diversity of natural product structures to a manageable limit, the scaffolds rather than entire molecules were classified and arranged hierarchically, that is, all rings, connecting aliphatic linkers, and ring-based double bonds. For each scaffold, a branch is generated by iterative deconstruction of one ring at a time, guided by a set of rules. The resulting smaller scaffold is termed the "parent" and the larger scaffold the "child". Repeated removal of rings by the algorithm as long as possible (usually until only one ring is left) generates a scaffold branch, in which each parent scaffold is the substructure of its child. In other words a child scaffold grows out of the parent scaffold. In a final step, all branches are merged to yield the final scaffold tree. The "natural product tree" has provided guidance for the design and synthesis of several compound collections inspired by natural products and to gain insight into new biology. However, since not only natural products but also numerous non-natural products, including in particular many drugs and agrochemical ingredients, are biologically relevant, it was necessary to extend the scaffold tree approach beyond natural

products. This required changing the set of rules that had been employed for the establishment of the tree, and which, for example, included the rule that required each parent scaffold to occur in natural products.

To this end, a new set of 13 rules[2b, 16] was developed, which introduced further features and guiding arguments. For example, it ensures that each child scaffold is connected to only one parent scaffold. Although information in terms of alternative branches is partly lost this way, the reduction is key to obtaining a treelike diagram logically and that is amenable to visual inspection rather than by an extended graph. Such simplification facilitates "scaling of human cognition",[17] that is, enabling the human mind to interact with and cope with large amounts of data. The choice of rules and their priorities in the devised set of rules is also guided by knowledge and experience of synthetic and medicinal chemistry, and, therefore, to some extent is subjective. The entire procedure yields a very flexible, yet intuitive classification that can accommodate virtually any molecule and connect it to others through substructure relationships.

The initial focus on the chemical space explored by nature in the first version of the natural product scaffold tree also produced many "holes" in the scaffold tree. These holes arose where structures were missing that had either not been generated through evolution or had yet to be discovered. These holes made tree construction and overlays of scaffold trees from different sets of molecules very difficult, if not impossible. The improved version of the scaffold tree set of rules also allows for and generates "virtual scaffolds" to complete the tree. Such scaffolds are not contained in and do not represent molecules in the original data set to be analyzed, but are derived from the iterative deconstruction and are generated in silico. These virtual scaffolds fill the gaps and provide clear opportunities for chemistry and biology research. "Brachiation", a term adopted from anthropology, which describes the movement of gibbons in botanical trees, was introduced to describe the movement along the branches of scaffold trees, from larger, more complex towards smaller and structurally less-complex scaffolds. During brachiation, the type of bioactivity is assumed to be retained, but may vary, for example, in terms of potency (Figure 4). Notably, in this
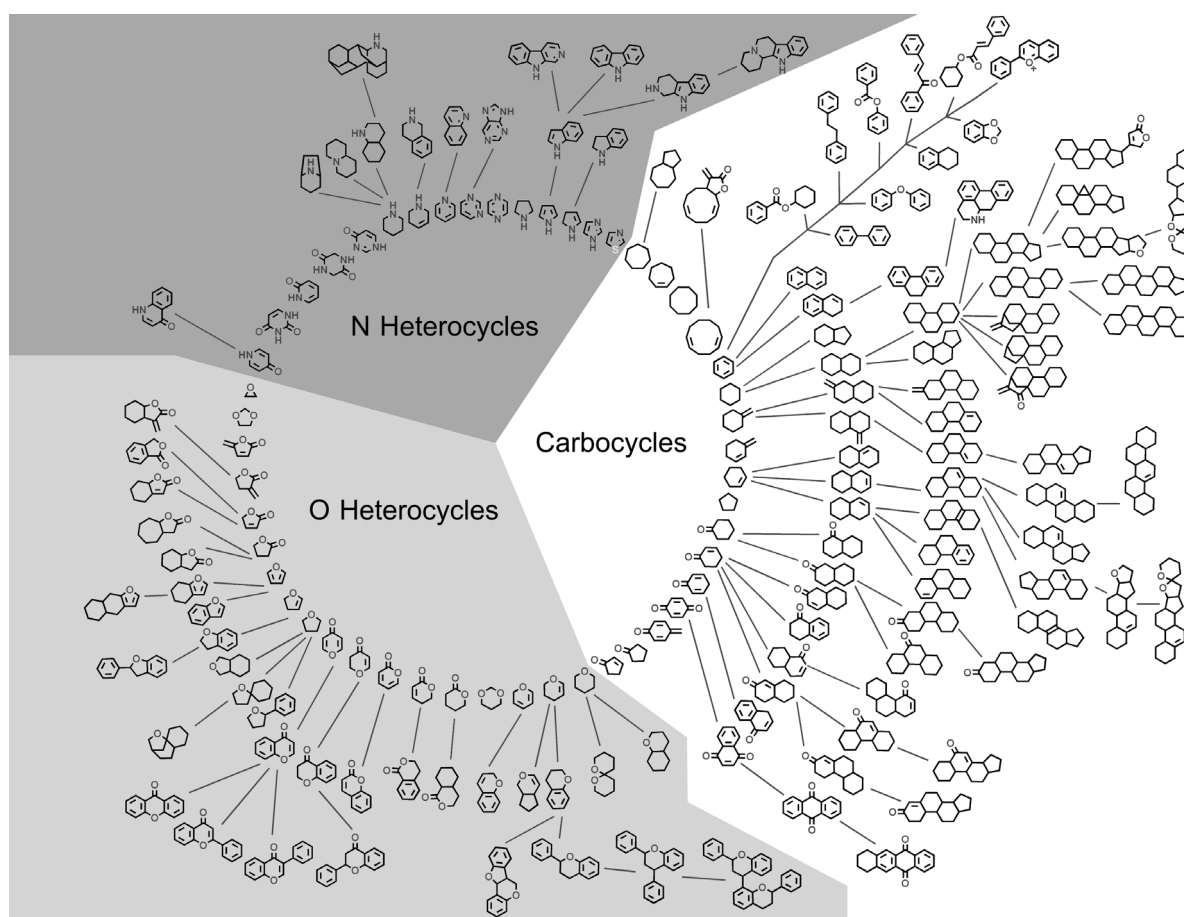
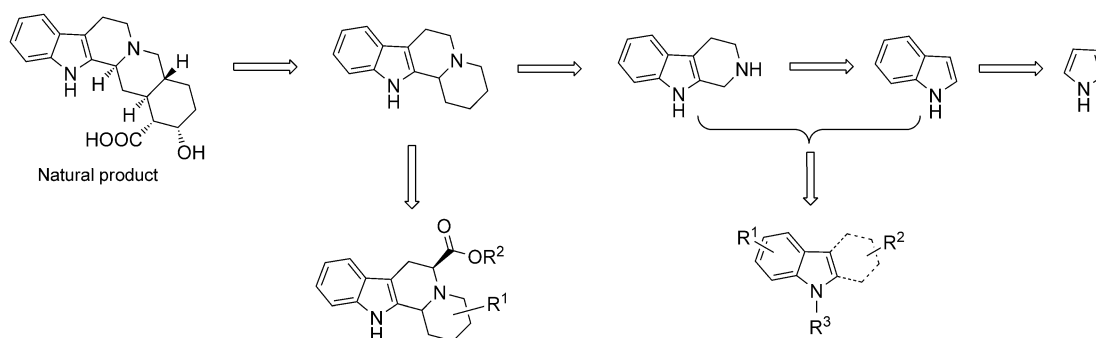**Figure 3.** The natural product scaffold tree.

Figure 4. Brachiation in scaffold tree branches exemplified by an example from the N-heterocyclic part of the natural product tree. In this case, brachiation leads from the pentacyclic scaffold of the alkaloid yohimbine, an inhibitor of the phosphatase Cdc25 A, to tetra-, tri-, and bicyclic scaffolds. Compound collections based on these ring systems were synthesized and yielded several inhibitors of Cdc25A (for details see Section 3.2).

approach, brachiation through scaffold structures of natural products proceeds along lines of biological relevance. Thus, it differs fundamentally from structure simplification based exclusively on chemical arguments, such as synthetic tractability of smaller scaffolds or retrosynthetic considerations. Brachiation is based on the assumption that smaller scaffolds share properties with the larger molecules into which they are

incorporated—the concept underlying fragment-based drug discovery.

Brachiation also inspired and suggested complementing and extending the strictly chemistry-based construction of the natural product tree to biology-guided scaffold trees.[18] For example, attempts to place morphine in the scaffold tree of natural products failed (see Figure 22), but complementation

of the tree with scaffolds of non-natural products having the same kind of bioactivity as morphine would have allowed us to close the gap. Biology-guided scaffold trees offer a view on chemical space from a different perspective, by employing bioactivity as a guiding criterion during branch construction. In a sense, they represent "brachiation"-based scaffold trees, that is, scaffold sequences with a particular kind of retained, but graded, bioactivity. In this bioactivity-guided navigation of chemical space, all possible parent–child scaffold pairs are generated for every given child scaffold in each deconstruction step. Branch construction is then guided by the same kind of bioactivity, for example, in vitro activity against a particular target. Multiple branches are constructed in cases where there are multiple parent–child pairs that exhibit similar biological activity. In a final step, the longest branch with the fewest gaps is selected. Combination of branches to form the scaffold tree is performed by analogy to the chemistry-guided scaffold trees.

The value of scaffold trees largely depends on annotation from various sources, including origin, frequency of occurrence, average biological activity, and target information. The need to visualize and intuitively use and interact with extensively annotated scaffold trees in a dynamic manner led to the development of a JAVA-based program named "Scaffold Hunter".[19] Scaffold Hunter facilitates the automatic visualization, filtering of and navigating through scaffold trees in an intuitive manner. It offers property- and structure-based filtering, a wide range of color-based highlighting methods, as well as a wide range of customizable settings. Interactive navigation in the scaffold trees includes zooming, panning, as well as automatic construction of subtrees consisting of selected scaffolds. A second program, ScaffoldTreeGenerator allows the generation of scaffold tree databases from SD files,[20] a format that can be exported from widely available structure sketching programs, including ChemDraw[21] and ISIS Draw,[22] and to import additional data including bioactivity values. Together, the two programs allow chemists and biologists—often non-experts in chemoinformatics—to generate, visualize, and analyze scaffold trees generated from virtually any set of chemical structures and to annotate with data. They are publicly available at scaffold-hunter.sourceforge.net.

The application of Scaffold Hunter and the chemistry- and biology-guided scaffold trees in the discovery of bioactive molecules depends on the data set to be analyzed and the guiding problem. Chemistry-guided scaffold trees can be constructed for any given set of molecules, irrespective of annotation. They can be used to merge different data sets and to guide synthesis efforts. Bioactivity-guided scaffold trees can incorporate large bioactivity data sets and guide prospective bioactivity annotation. In many cases, the two approaches will be complementary and, if the data allow it, should be explored in parallel.

A particularly promising application is the exploration of virtual scaffolds in chemistry-guided scaffold trees. Since virtual scaffolds should share properties with their neighboring scaffolds, they should be good templates for the design of compound collections enriched with biological activity. A similar scenario should be valid for biology-guided scaffold

trees. Scaffolds representing gaps in the branches, that is, molecule classes without annotation for the target of interest, may be good starting points for the development of compound collections with a particular expected activity.[19]

To investigate whether virtual scaffolds filling the gaps in chemistry-based scaffold trees represent promising starting points for compound synthesis, a scaffold tree from 765 135 ring-containing structures in PubChem, for which biological or biochemical assay data were available, was generated.[19] The target proteins given in PubChem were then compared with targets listed in WOMBAT,[23] a database assembled from molecules and their bioactivity data in the scientific literature. Promising virtual scaffolds that were next to scaffolds annotated with activity were identified from targets present in both databases. WOMBAT was then searched for compounds containing the virtual scaffold and active against the same molecular target, thus filling the gaps in the Pubchem dataset.

The potential of the approach for the prospective identification of promising scaffolds was demonstrated by analyzing the pyruvate kinase screen data set deposited in PubChem.[24] Four scaffolds were selected to assemble a small compound collection, which was analyzed biochemically for pyruvate kinase inhibition or activation. Nine compounds displayed an AC_{50} value of $< 10\,\mu M$ in the screen. Virtual scaffolds in branches with inhibitory activity yielded six inhibitors, and virtual scaffolds in activator branches yielded three activators. A search in *Chemical Abstracts* found that none of the compounds had been linked to any kind of kinase inhibiting activity before (Figure 5).[25]

As indicated above, bioactivity-guided scaffold trees can be applied in a similar manner. Such an analysis of the WOMBAT database with bioactivity-guided scaffold trees revealed that brachiation is possible for 1/3 of all targets, and yielded numerous cases where brachiation covers 3–9 steps. Among these targets are members of all major classes of drug targets, that is, kinases, opiod receptors, G-protein-coupled receptors (GPCRs), and enzymes (Figure 6). Brachiation was found to be more common than expected, and to be the rule rather than the exception.

Two sequences targeting 5-lipoxygenase (5-LOX) and estrogen receptor alpha (ERα) were probed by means of biochemical assays to assess the potential exploitation of gaps identified in the branches. The branch for 5-LOX contained compounds containing one to seven rings, with an annotation gap at compounds with three rings (Figure 7). The ERα branch spans from compounds with six rings to only one ring, with a gap at the bicyclic scaffold (Figure 8). Of four compounds designed based on the tricyclic 5-LOX scaffold, two showed single digit micromolar IC_{50} values in a cell-based assay system. For ERα, eight molecules were designed on the basis of the identified bicyclic scaffold. Concentration-dependent measurements with a fluorescence-based assay yielded one inhibitor with an IC_{50} value of $20\,\mu M$ for ERα and $4.6\,\mu M$ for ERβ. Whereas the potency may seem limited at first glance, a closer inspection shows that the unoptimized inhibitor has a potency of only about 100-fold less than the natural substrate estradiol.
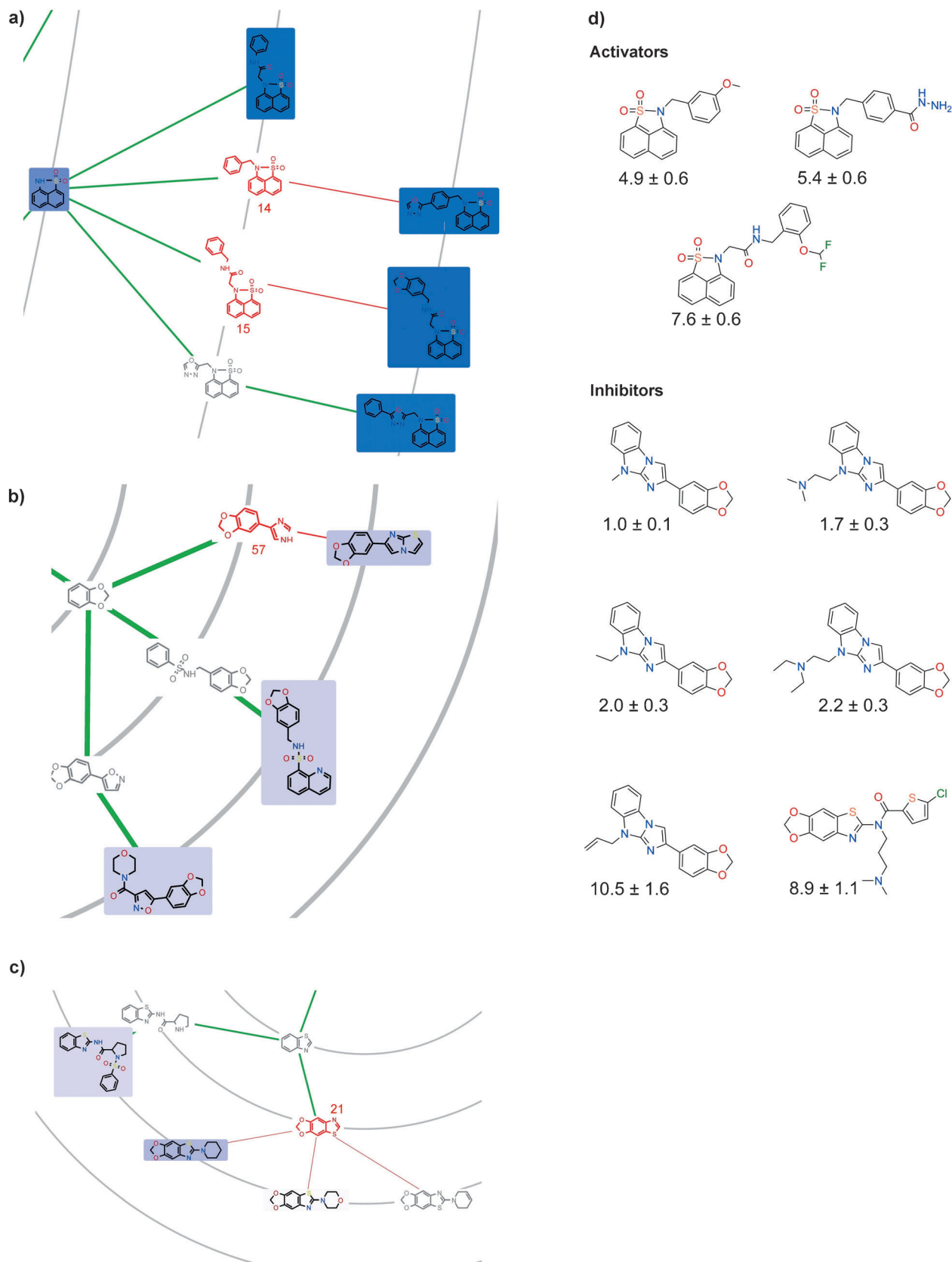
Figure 5. Selected branches of the scaffold tree derived from the pyruvase kinase screening data set and results of the screens. The four virtual scaffolds selected are shown in red, together with the corresponding number of compounds containing each scaffold. a) This branch consists of several scaffolds that are good activators of pyruvate kinase. b,c) Branches that represent inhibitors of pyruvate kinase. Additional virtual scaffolds are shown in gray. The blue shading highlights the mean $\log(AC_{50})$ values obtained from the data set (darker shading represents higher activity). The images were exported from Scaffold Hunter. d) Inhibitors and activators of pyruvate kinase with $IC_{50} \leq 10$ μm from the pyruvate kinase screen (data available on PubChem, assay ID 2941). Reproduced from *Nat. Chem. Biol.* **2008**, *5*, 581–83.
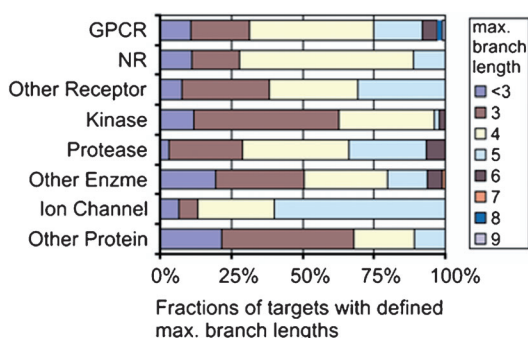
Figure 6. Brachiation length, that is, the number of rings that can be removed from the scaffold while retaining similar bioactivity for the most important target classes. The distribution of the lengths of the longest branches per target over the target classes reveals that for most of the target classes more than half of the targets had branch lengths of 4 or more.

These findings indicate that Scaffold Hunter facilitates the identification of gap-filling scaffolds in chemistry-guided as well as bioactivity-guided scaffold trees. These structures represent promising starting points for the design of focused collections of small molecules with biological relevance for the target of interest.

Brachiation can be a viable strategy to identify structurally simple analogues, in particular in the design of libraries inspired by natural products. However, in many cases, hardly any knowledge about the bioactivity profile of the natural products is available. Therefore, methods to prospectively annotate bioactivity would be invaluable. The observation that brachiation is a widespread phenomenon and that scaffold classes occupying gaps in (non-)annotated scaffold trees may share bioactivity with their neighboring scaffolds suggested that biological annotation can be inferred from an annotated to a non-annotated set of molecules by merging the scaffold trees derived from both sets (Figure 9). Besides the direct annotation of scaffolds present in both data sets, the annotation should propagate along the branches of the scaffold tree by analogy to brachiation. Thus, a much broader annotation can be achieved, as even scaffolds present only in the annotated data set can pass on their annotation to neighboring scaffolds in the same branch.

This hypothesis was explored by using the bioactivity information in the WOMBAT database to annotate the natural product structures in the γ-pyrone branch of the Dictionary of Natural Products (DNP). This led to the merging of the respective scaffold trees derived from DNP and WOMBAT.[26]

Several scaffolds were identified from WOMBAT where activity spanned more than two out of five scaffolds in the branch. A compound collection with 500 γ-pyrones spanning five scaffolds in three hierarchy levels of the scaffold tree was assembled. This library was analyzed for inhibition of monoamine oxidases A and B, the signal transducers and activators of transcription (STAT) proteins STAT1, STAT3, and STAT5b, as well as acid sphingomyelinase, as annotated in WOMBAT. Notably, inhibitors were found for all proteins (Figure 10), which in some cases were (isoenzyme) selective and similar to structures independently identified by unbiased screening efforts.[27]

These findings demonstrate that scaffold trees can be used favorably to identify novel scaffolds for the development of compound libraries by filling gaps within a given data set. Furthermore, they open up a possibility to prospectively annotate bioactivity for non-annotated compound classes by merging scaffold trees. Prospective annotation is of particular relevance for predicting the bioactivity of natural product classes. However, the approach is restricted to the scaffold level, and does not necessarily include bioactivity annotation of individual natural products.
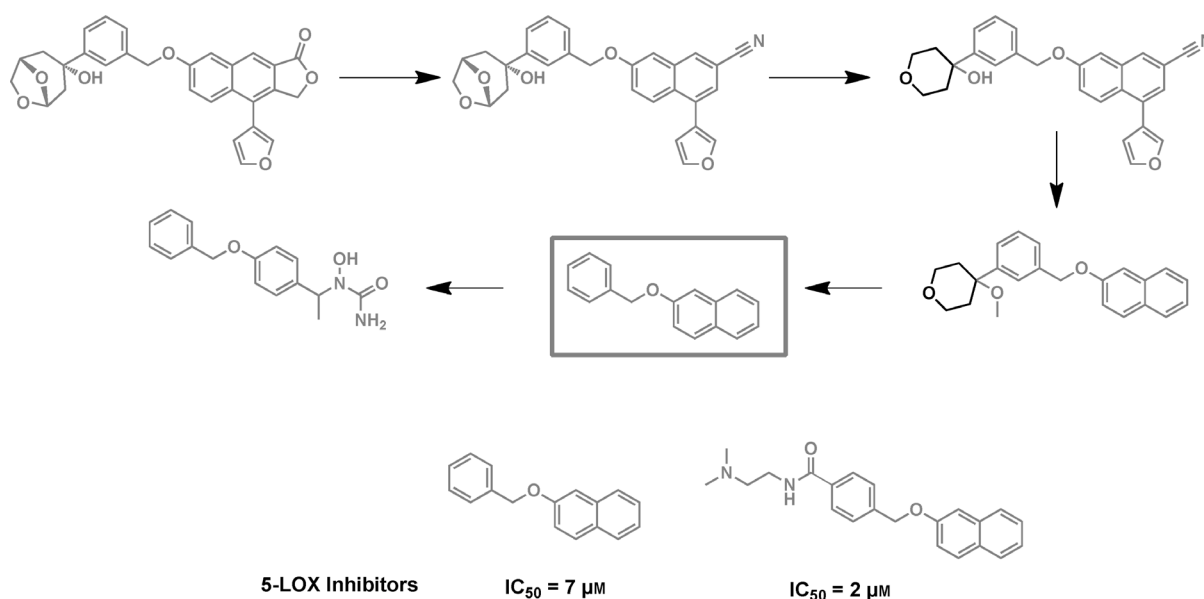

Figure 7. Biology-guided scaffold branch of 5-LOX inhibitors. The scaffold in the box has no activity annotation for 5-LOX and served as the template for a small collection of compounds, two of which were micromolar inhibitors.
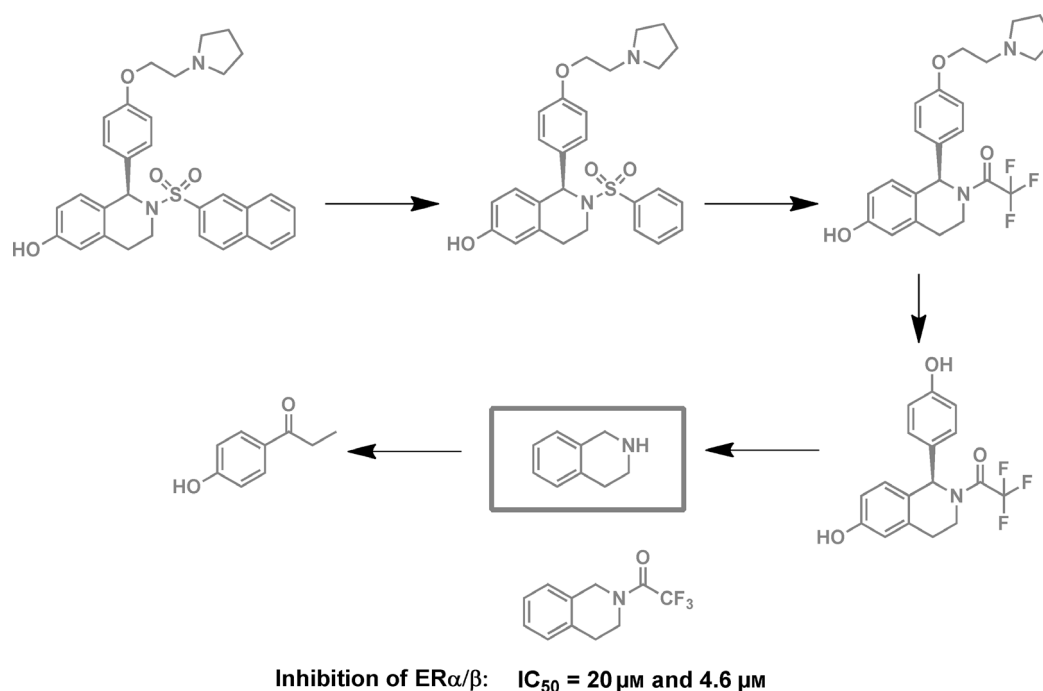
Inhibition of ERα/β: IC$_{50}$ = 20 μM and 4.6 μM

Figure 8. Biology-guided scaffold branch of estrogen receptor α/β (ERα/β) inhibitors. No compounds incorporating the scaffold in the box were known that modulate ERα/β activity. Hence, the scaffold served as a template for a small collection of compounds and yielded one inhibitor.

2.2. Protein Structure Similarity Clustering (PSSC)

Structural complementarity between a small molecule and a protein binding site is required for productive molecular interactions, which usually involve the substituents of the small molecule and the side chains of amino acids embedded in the protein. High sequence similarity usually leads to high structural similarity and, hence, also to the binding of similar ligands. The definition of complementarity in the PSSC concept extends beyond sequence similarity towards structures with low sequence similarity but still a high structure similarity. Whereas high sequence similarity can be identified by sequence homology analysis, the structural similarity requires a structure-based method.

The three-dimensional arrangement of interaction points in space is determined by the scaffold, that is, the molecular framework or the backbone arrangement (=subfold) in the ligand-sensing binding site of the protein (Figure 2). Hence, complementarity at the scaffold level, although more abstract, should also be required. Thus, ligands with similar scaffolds will be bound by proteins with similar subfolds in the binding site. This hypothesis defines the basic reasoning of protein structure similarity clustering (PSSC), which groups proteins according to the structural similarity of their binding-site subfolds (Figure 2). These clusters can then be exploited, for example, to identify promising types of small-molecule structures for proteins or to find potential alternative targets of a given compound class.

One example for the potential use of PSSC was identified from literature data.[28] The farnesoid X receptor (FXR), a nuclear hormone receptor which plays a key role in metabolic

diseases, was clustered together with the estrogen receptor β (ERβ)[29] and the peroxisome proliferation-activated receptor γ (PPARγ).[30] Although the proteins exhibit sequence similarities below 20%, they share a highly conserved subfold around the binding site (Figure 11). A subsequent literature search identified the natural product genistein, which is a known inhibitor of ERβ and PPARγ. The drug troglitazone based on the same scaffold is a known PPARγ modulator. Notably, hits from a screening of a library of 10000 benzopyrans for FXR inhibitors could have been predicted by PSSC. In addition, the benzopyran library also yielded ligands for other members of the PSSC, thereby further supporting the application of PSSC in library design.

Initially, protein structure similarity clusters were defined on the basis of structural similarity searches performed using the FSSP online database, which is based on fold comparisons of a nonredundant subset of the Protein Databank (PDB) using the DALI alignment program. The resulting search list was then analyzed according to similarity and for interesting results, for example, alignments with high structural but low sequence similarity. In a final validation step, the ligand-sensing cores of the cluster members—spherical cutouts of the protein structure centered on the binding site—were manually extracted and their structural similarity was visually assessed. However, this procedure also led to false-positive alignments where regions remote to the binding site aligned well and gave a misleading score. Thus, it was improved[31] by first automatically extracting the ligand-sensing core, that is, the subfold surrounding the binding site. This "ligand-sensing core" was then submitted to a structural similarity search against the FSSP database, thereby ensuring alignment with the binding site of interest. This step drastically decreased the number of false-positive hits and allowed more focused follow-up investigations (Figure 12).

The initial version of the PSSC approach was applied in the identification of novel inhibitor chemotypes. In this approach, a scaffold from a known natural product inhibitor of the phosphatase Cdc25A was used to identify novel inhibitors of other proteins clustered with Cdc25A in a PSSC, for example, acetylcholine esterase (AChE) and 11β-hydroxysteroid dehydrogenase 1 (HSD1, Figure 13). Thus, a
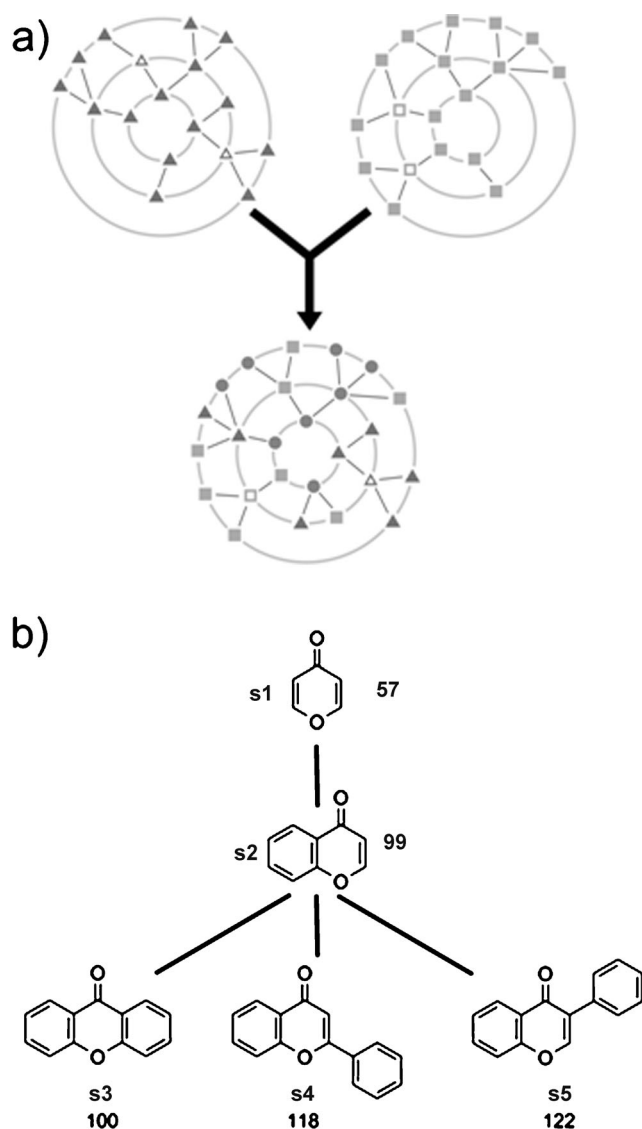
Figure 10. Target classes annotated in the γ-pyrone branch for which novel inhibitors were found. Interestingly, hits were found for all target classes with notable potency and selectivity, given that these are unoptimized compounds.

Figure 9. a) The merging of two scaffold trees (triangles and squares) creates a new tree. In this new scaffold tree, nodes can represent molecules either from both trees (filled circles), from one tree only (filled trangles/squares), or from neither tree (outlined triangles/ squares). Annotation, for example, about target proteins, can be directly transferred if a node represents molecules from both trees or indirectly through brachiation. b) The γ-pyrone library was comprised of 500 molecules spanning five different scaffold types in the tree.

library of hydroxybutenolides important for inhibition of the phosphatases yielded novel inhibitors for AChE and the 11βHSDs (for a more-detailed discussion see Section 3.2).[32]

In a recent example, a protein structure similarity cluster was constructed based on gastric lipase, an enzyme modulated by compounds containing a β-lactone structural motif, including the marketed drug tetrahydrolipstatin (Orlistat). A similarity search with the ligand-sensing core of gastric lipase yielded a list of structurally similar proteins, including acylprotein thioesterase 1 (APT1). A collection was synthesized based on tetrahydrolipstatin and was biochemically tested against the thioesterase. Notably, several inhibitors of
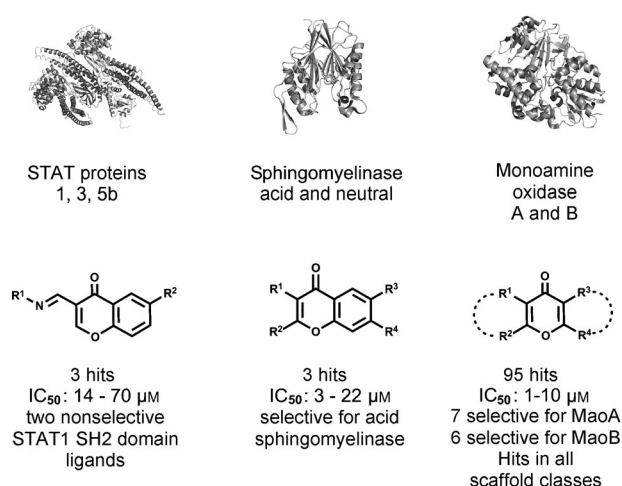
APT1 with IC_{50} values in the low micromolar and nanomolar range were discovered (for a detailed discussion see Section 3.2).[33]

These two successful applications of PSSC indicate that the method may indeed provide a viable route to identify target–ligand pairs. However, the limited number of studies currently completed means that conclusions about the general applicability of PSSC or about the scope of the method would be premature.[34]

3. Biology-Oriented Synthesis (BIOS)

The structural classification of natural products (SCONP) and its extension to non-natural products and PSSC provide two complementary approaches for the identification of biologically relevant compound classes in vast chemical space. Either applied alone or in a synergistic way, they define the underlying reasoning of an approach we term biology-oriented synthesis (BIOS; Figure 14).[5c,9b,35]

In BIOS, biological relevance is the prime criterion for the selection of compound classes and scaffolds that inspire the synthesis of compound collections enriched in bioactivity. BIOS-based compound libraries are typically not and do not have to be large. In our experience, screening of such libraries yields initial hits with rates of 0.2–1.5%, thereby calling for library sizes of 200–500 compounds to initiate further development. Their synthesis, however, may require the application of elaborate chemistry methods and demanding multistep sequences, in particular if libraries inspired by natural products have to be synthesized. However, this investment in chemical development is well-balanced by the smaller library size needed. In a sense, BIOS offers relevant compounds, but demands more of chemistry.

The reduction in structural complexity compared to the guiding natural products may result in the initial hits obtained from screening the primary BIOS libraries in biochemical and
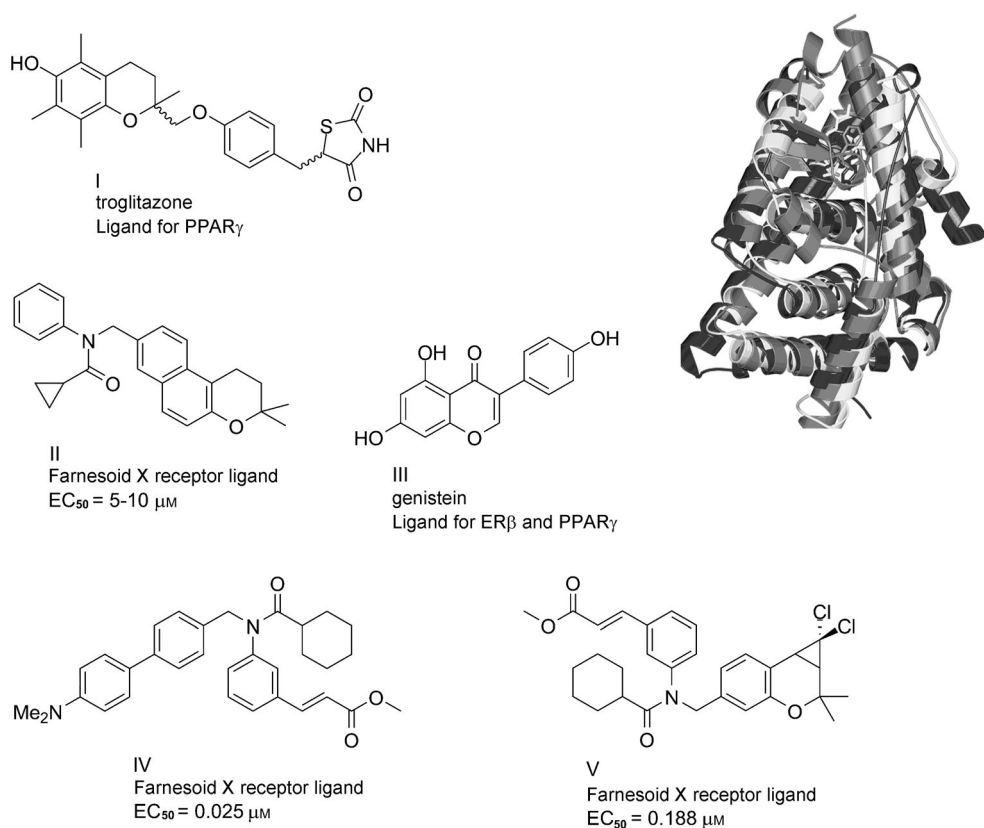
Figure 11. Protein structure similarity cluster of ERβ complexed with genistein (III, dark gray), PPARγ with rosiglitazone (medium gray), and FXR (light gray). The overlay of the ligand-sensing core structures illustrates the structural similarity of the structures of the benzopyran-based ligands for ERβ, PPARγ, and FXR.

I
troglitazone
Ligand for PPARγ

II
Farnesoid X receptor ligand
EC_{50} = 5–10 μM

III
genistein
Ligand for ERβ and PPARγ

IV
Farnesoid X receptor ligand
EC_{50} = 0.025 μM

V
Farnesoid X receptor ligand
EC_{50} = 0.188 μM

Figure 12. Revised PSSC procedure. The false-positive rate is drastically decreased as ligand-sensing cores are extracted before the alignment, thereby focusing on the structural similarity in the relevant part around the binding site.

biological assays being nonselective. Furthermore, they may also simultaneously target several proteins with similar ligand-sensing cores, and may be only of limited potency, for example, showing IC_{50} values in the micromolar range. At such concentrations, possible promiscuous binding also has to be considered in the screens, which requires careful development of screening conditions and follow-up experiments, including appropriate control experiments. However, these are frequently encountered problems in the screening for and development of both "tool compounds" and drug candidates in general. They call for further elaboration of initial hits to generate potent and selective "tool compounds",[2a] which is the day-to-day work of the medicinal chemist and often the chemical biologist in any case.

BIOS was originally developed on the basis of an analysis of natural product structure. However, the key criterion in

BIOS is biological relevance, not occurrence in nature. Hence, BIOS includes, but is not restricted to, natural products but instead extends well into the chemical space of non-natural products. Notably, this extension includes the numerous non-natural compound classes that were investigated in more than 100 years of pharmaceutical development.

Both SCONP- and PSSC-based compound libraries are based on structural considerations. In SCONP-derived libraries, the scaffolds of bioactive small molecules guide the design and synthesis efforts, and biological relevance is delineated from the biosynthetic origin and biological activity. In PSSC-derived libraries, collective protein structure fuels the reasoning and provides the basis of the biological relevance of the designed compound collections. Both approaches on their own suffice to inspire BIOS and may serve as hypothesis-generating methods.

Besides their individual application, SCONP and PSSC may be applied synergistically and reinforce each other. The development of 11βHSD1 and of APT1 inhibitors, mentioned briefly above and discussed in more detail below, are representative examples, which convincingly demonstrate the power inherent to this approach.

The synergistic approach is often hampered by a lack of data for bioactivity annotation, especially for natural products and by the lack of protein crystal structures, in particular with bound ligands or inhibitors. This lack of protein structure data is strikingly apparent if biosynthetic arguments are employed to invigorate the BIOS approach. In principle, similarity between the structures of proteins involved in the biosynthesis of classes of natural products and other proteins should indicate potential targets of natural products.[36] Accordingly, comparison of the protein fold topology (PFT) of the enzymes chalcone synthase (CHS), chalcone isomerase (CHI), and anthocyanidin synthase (ANS), which catalyze key steps in the biosynthesis of naturally occurring chalcones and flavanoids, indicated a similarity of the catalytic sites of these enzymes with the active site of phosphoinositide-3-kinase (PI3K). Although CHS, CHI, ANS, and PI3K are considered
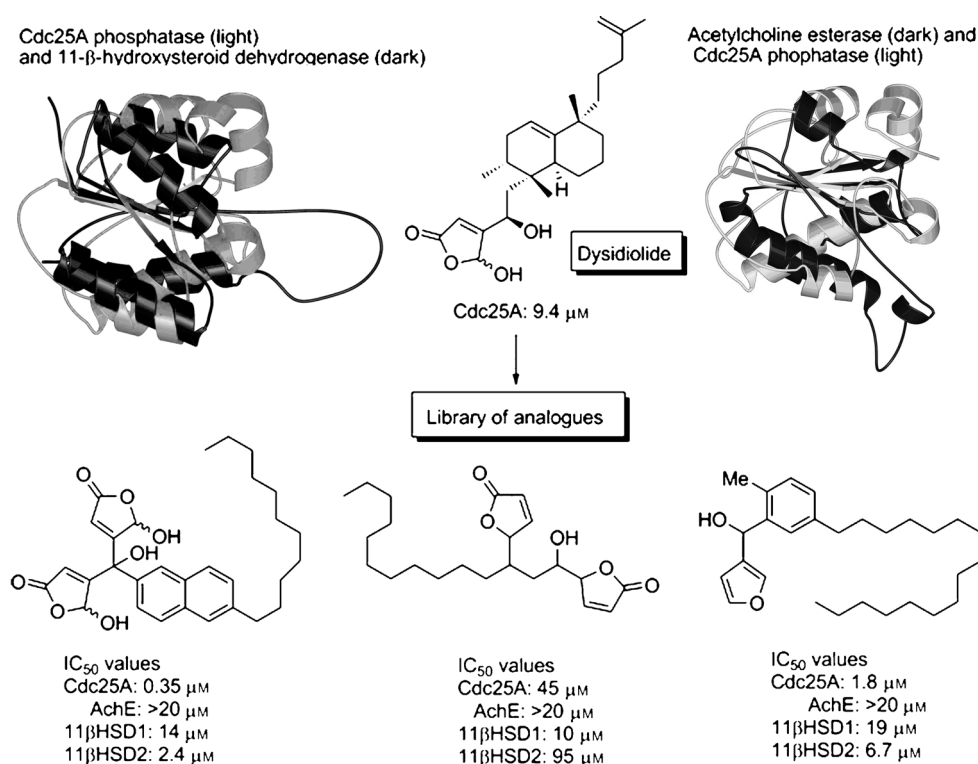
Figure 13. Cluster of the HSD, Cdc25a, and AChE proteins and the corresponding hit compounds from a dysidiolide-inspired compound collection.

inhibitors to be identified.[37] Although this approach establishes the link between biosynthetic enzymes of a natural product and potential targets, a lack of knowledge of many biosynthetic enzymes and their structures limits the application of this approach.

3.1. BIOS in the Development of Compound Collections

Exploring the BIOS concept for medicinal chemistry and chemical biology research requires the synthesis of compound collections based on biologically relevant structural frameworks. Natural products represent a major source of bioactive molecules. However, the limited accessibility of these molecules from natural sources and/or by synthetic or semi-synthetic methods often

very different on the basis of their fold classification, similar arrangements of different secondary structures, namely PFT, were observed. Indeed, chalcones were among the first kinase

limits their further exploration in the biological sciences. This generates the need to synthesize complex natural product like molecules in sufficient amounts and numbers,
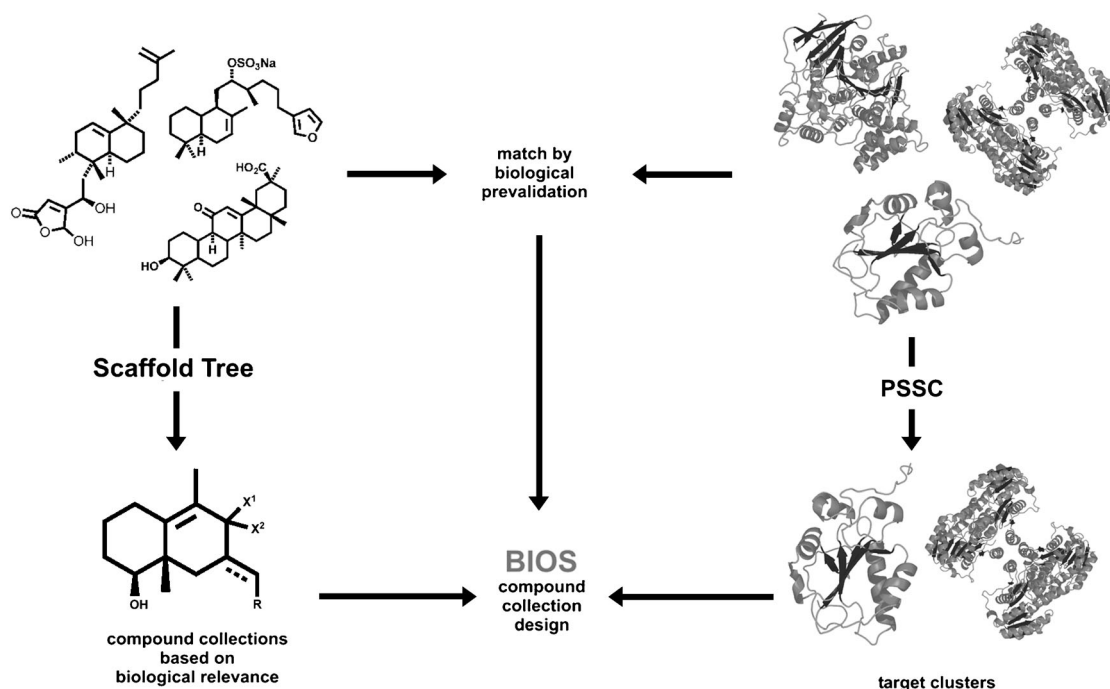

Figure 14. BIOS connects chemical and biological space, that is, protein structure similarity clusters and small-molecule compound collections through biological prevalidation. This extends well-beyond natural products and includes all compounds with known biological relevance.

and calls for the development of new strategies and methods amenable to the formats of compound library synthesis. A synergistic approach that utilizes the power of contemporary organic synthesis and the technology of combinatorial and parallel synthesis is required to synthesize focused libraries based on the core frameworks of natural products and other biologically relevant chemotypes. Chemical transformations and reaction sequences (with respect to overall high yields and reduced number of individual reactions steps) that utilize readily accessible substrates to provide complex molecular architectures based on natural products are highly desired and challenging. This challenge has been met, for example, by recent developments of multicomponent reactions, cascade and domino reaction sequences, one-pot multicatalytic reactions and asymmetric solid-phase syntheses that have led to natural product inspired molecules.[38] Given the diverse ring systems and core scaffolds present in natural products, the choice of preferred ring systems as targets for library synthesis is often not clear. Statistical analysis of the scaffolds of the natural products in the DNP revealed that more than half of the small natural molecules under 1000 g mol^{-1} in the database contain two, three, or four rings. This indicates that systems with two to four rings provide good starting points for the development of compound collections inspired by natural products (Figure 15). In the synthesis it should be considered
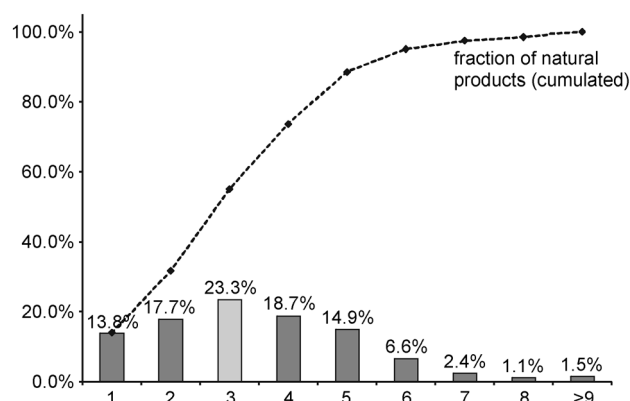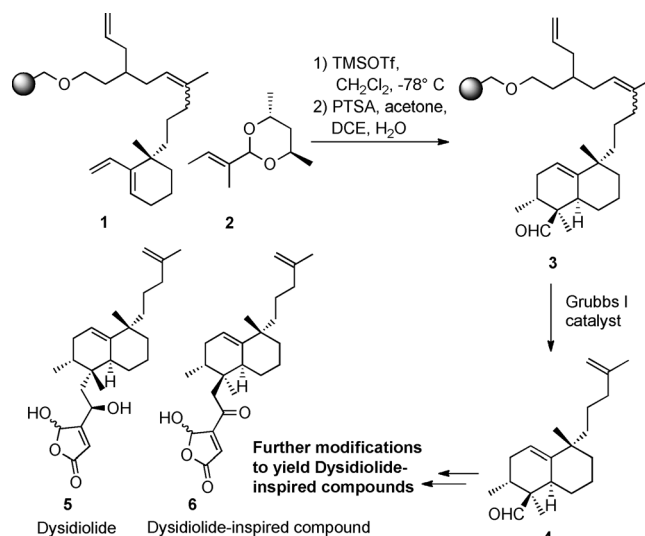
analogues of the natural phosphatase inhibitor dysidiolide (**5**; Scheme 1). To this end, chiral dienophile **2** was employed to enhance the stereodirecting influence of the resin-bound


Scheme 1. Solid-phase synthesis of dysidiolide-inspired compounds. DCE = 1,2-dichloroethane, Tf = trifluoromethanesulfonyl, PTSA = *p*-toluenesulfonic acid, TMS = trimethylsilyl.

Figure 15. Occurrence of scaffolds with different numbers of rings in natural products. 20.8% of all natural products contain three rings and mark the maximum of the distribution. However, the number of scaffolds with two or four rings lie within one standard deviation, such that 52.8% of all natural products contain two, three, or four rings.

that unlike collections derived from natural products, in which the scaffold is identical to the backbone of a given natural product, in collections inspired by natural products, the scaffold may not be identical but closely related to the guiding natural product itself. The scaffolds will typically be constructed by de novo synthesis, thereby allowing the introduction of substituents and variation of the substituent pattern and stereochemistry (see Figure 1).[14] Below we summarize selected syntheses of compound collections inspired by natural products. For recent overviews of the field the reader is referred to more comprehensive reviews.[39]

A natural product inspired synthesis of dysidiolide-like molecules was developed to identify biologically active

chiral diene **1**.[40] The bicyclic scaffold **3** was built up by Diels–Alder reaction of diene **1** with acetal **2** derived from tiglic aldehyde and displayed an *endo/exo* ratio of 91:1 and a selectivity of 95:5 in favor of the desired *endo* isomer. The cycloadduct **4** was released by a ring-closing metathesis (RCM) reaction. Further modifications of the cycloadduct **4** provided analogues of dysidiolide. Biological evaluation of this focused small library revealed inhibitors of phosphatases and cytotoxic activity against different cancer cell lines, with dysidiolide-like molecule **6** being the most potent inhibitor of the phosphatase Cdc25C with an IC$_{50}$ value of 0.8 μM.

The synthesis of a compound collection, particularly on a solid phase, often requires adaptation of known chemical transformations to a format for library synthesis. For example, developments in solid-phase asymmetric synthesis have facilitated the generation of natural product inspired collections.[38b] A prominent example is the use of enantioselective carbonyl allylation—one of the most important functional group transformations—for the stereoselective solid-phase synthesis of a collection of natural product inspired δ-lactones (Scheme 2).[41] The synthesis design included multiple stereocomplementary allylation reactions on the polymeric carrier followed by a ring-closing metathesis to provide natural product analogues (Scheme 2). Therefore, prior to the synthesis of the library, reaction conditions for the highly enantioselective and high-yielding allylation of an immobilized aldehyde were identified by using *B*-allyl(diisopinocampheyl)borane (Ipc$_2$BAll) under different conditions. The allylation of the polymer-bound aldehyde **7** using L-Ipc$_2$BAll yielded resin-bound **8** in a *syn/anti* ratio of 85:15. Careful ozonolysis of the double bond yielded aldehyde **10**, which was subjected to a second allylation with L-Ipc$_2$BAll, and the
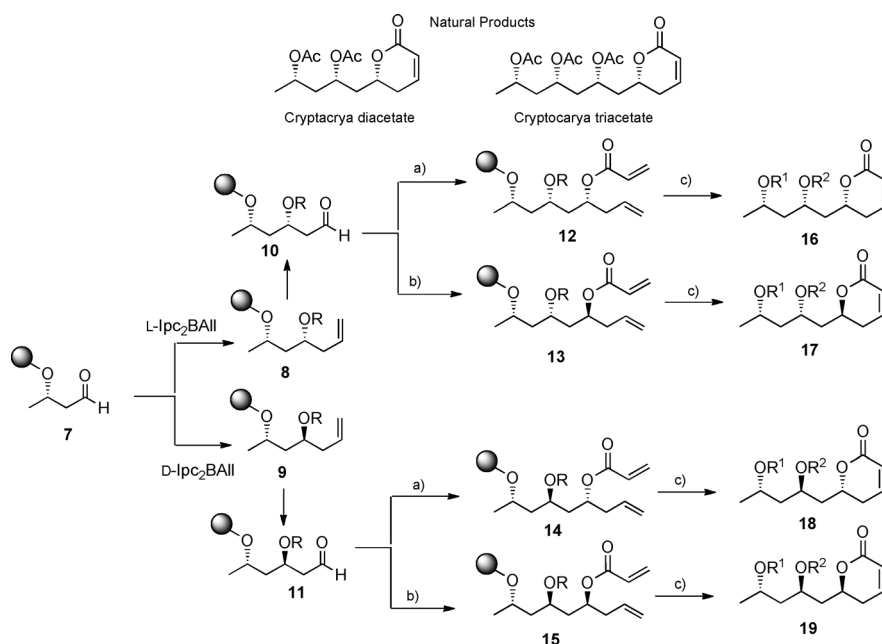
Scheme 2. Synthesis of stereoisomeric δ-lactones by using solid-phase asymmetric allylation of aldehydes as the key transformation. a) 1. L-Ipc₂BAll, 2. acryloylation; b) 1. D-Ipc₂BAll, 2. acryloylation; c) 1. Grubbs 2nd generation catalyst, 2. release from the resin.
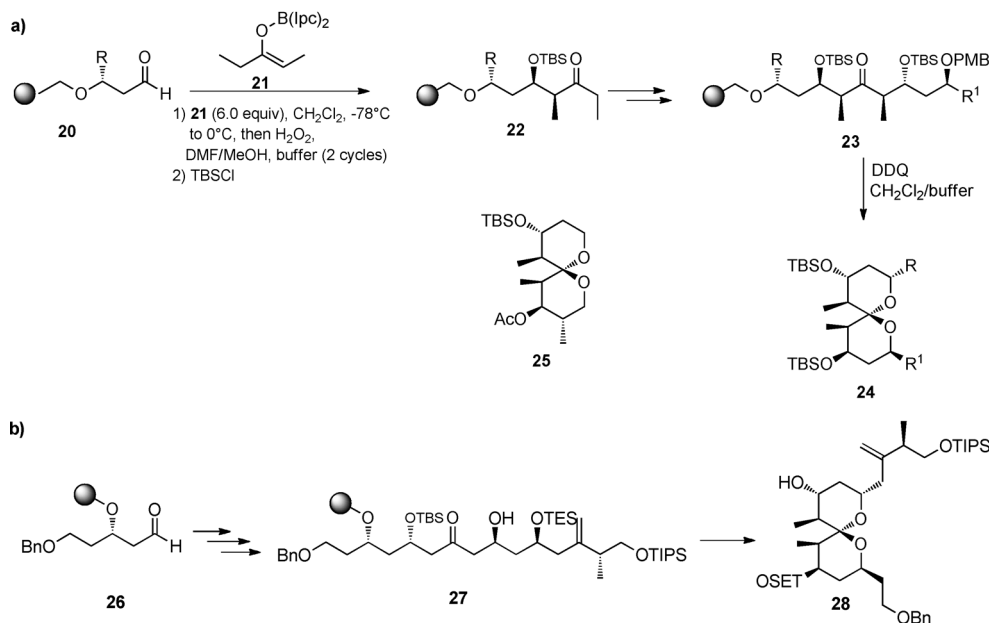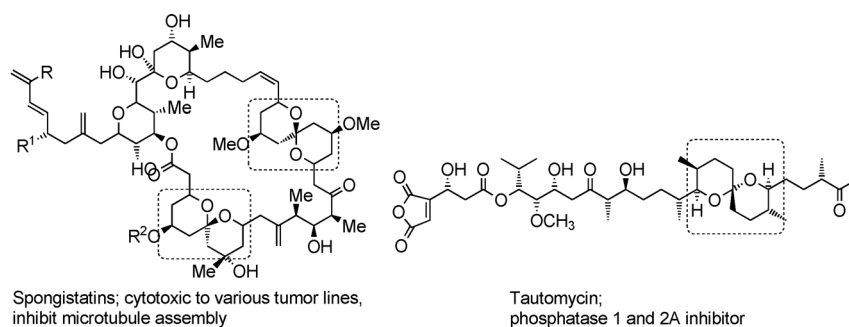
Scheme 3. Synthesis of natural product inspired spiroacetals. TBS = *tert*-butyldimethylsilyl, DDQ = 2,3-dichloro-5,6-dicyano-1,4-benzoquinone, TIPS = triisopropylsilyl, PMB = *p*-methoxybenzyl, TES = triethylsilyl, Bn = benzyl.

formed secondary alcohol was converted into acrylic acid ester **12**. Ring-closing metathesis with the Grubbs II catalyst provided the α,β-unsaturated lactone **16**, which was released from the polymeric support (with trifluoroacetic acid, TFA) and acetylated.

The all-*syn* isomer of cryptocarya diacetate was isolated in 11% overall yield after 11 steps by means of simple flash chromatography. This reaction sequence enabled all eight possible stereoisomers of the δ-lactone to be generated by carrying out the allylation reactions in a stereocomplementary fashion. Adapting an established asymmetric organic synthesis to the solid phase is often not straightforward, but the example illustrated above proves that existing synthesis methods allow, in principle, the generation of all stereoisomers of a given natural product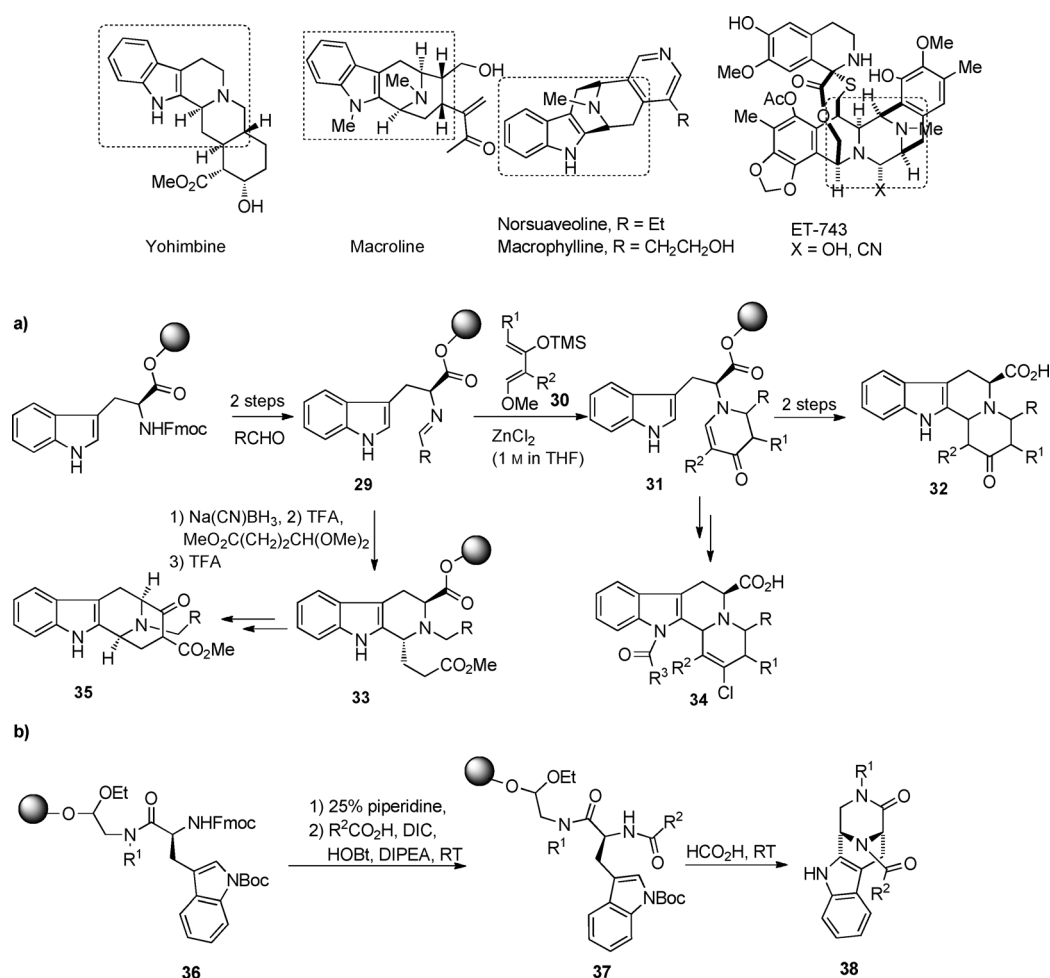. Among other examples involving asymmetric solid-phase synthesis, stereocontrolled aldol reactions on a solid phase were explored to create natural product inspired compound collections of spiroacetals. Natural products with spiroacetal structures occur widely in nature, and are known to have diverse biological activities.[42] In particular, the rigid spiro[5.5]ketal ring system is a fragment of various complex natural products that display a wide range of biological activities (Scheme 3). For example, the extraordinarily potent spongistatins, which inhibit tubulin polymerization, and the protein phosphatase inhibitor tautomycin[43] contain spiroacetal fragments within their macrocyclic frameworks. A natural product inspired synthesis of spiroacetals on a solid phase, with asymmetric aldol reactions used as the key transformations, was developed to identify the biological activities associated with the spiroacetal core, including bioactivity similar to the parent natural product. To

target spiro[5.5]ketals, an aldol reaction with resin-bound aldehyde **20** was performed with the preformed *Z*-boron enolate **21** to yield an enantio-enriched aldol adduct **22**. However, unlike in the solution-phase synthesis, the aldol reaction on a solid phase required two cycles with six equivalents of the chiral reagent **21** to achieve complete conversion of the aldehyde. Another *anti*-selective aldol reaction with the *E*-boron enolate on a solid phase built up the protected bis-β-hydroxyketones **23**, which are advanced precursors of the final spiroacetals **24**. Simultaneous cleavage of the PMB group and acetalization were achieved by oxidative cleavage with DDQ, thus releasing the spiroketals **24**. The diastereomeric ratios of the products revealed that the matched cases in the second aldol reaction yielded one diastereomer of spiroacetal **24** exclusively, whereas mismatched cases proceeded with lower stereoselectivity. Spiroacetal **25** (Scheme 3) of this collection was found to be an inhibitor of the phosphatases VHR and PTP1b, with IC$_{50}$ values of 6 and 39 μM, respectively. In addition, compound **25** distorted the correct organization of the microtubuli network in a human carcinoma cell line.[44]

In a similar approach, a fragment of the natural product spongistatin with the core spiroketal structure **28** was synthesized on a solid phase.[45] To this end, an immobilized β-hydroxy aldehyde **26** was subjected to two consecutive stereoselective aldol reactions to yield bis-β-hydroxy ketone **27**. Cleavage of the protected polyol **27** from the resin and in situ cyclization provided the spiroacetal **28**.

These examples illustrate how BIOS may allow the identification of structurally simpler starting points for library design while providing new classes of inhibitors. In an attempt to further explore natural product chemical space with the BIOS approach, indole alkaloid scaffolds were targeted. This was based on the finding that the structurally complex alkaloids yohimbine and ajmalicine are inhibitors of the protein phosphatase

Cdc25A. SCONP analysis and brachiation along the line of prevalidation given by nature led to tetracyclic indolo[2,3-*a*]quinolizidines (Figure 4). A solid-phase synthesis targeting indolo[2,3-*a*]quinolizidines yielded 450 compounds by means of a six- to eight-step synthesis sequence. The synthesis design involved the Lewis acid mediated Mannich–Michael reaction between immobilized D- or L-tryptophan imines **29** and electron-rich silyloxy dienes **30**. The enaminone products **31** were subsequently cyclized by treatment with acid or phosgene to yield tetracyclic ketones and vinyl chlorides. Further derivatization and base- or acid-mediated release from the polymeric carrier provided indoloquinolizidines **32** and **34** (Scheme 4) in high overall yield.[46] The collection of indoloquinolizine compounds contained two Cdc25A inhibitors with IC$_{50}$ values comparable to those of the natural products. The tryptophan imines **29** were also used to synthesize a macroline-inspired compound collection[47] consisting of tetracyclic indole derivatives **35** with a common cycloocta[*b*]indole framework. Thus, reduction of imine **29** followed by a Pictet–Spengler reaction with methyl-4,4-dimethoxybutyrate yielded the 1,3-*trans*-β-carbolines **33**. The necessary 1,3-*cis* arrangement to generate the tetracyclic

Yohimbine Macroline Norsuaveoline, R = Et Macrophylline, R = CH$_2$CH$_2$OH ET-743 X = OH, CN

Scheme 4. Polycyclic alkaloid inspired syntheses of compound collections. Fmoc = 9-fluorenylmethoxycarbonyl, DIC = diisopropylcarbodiimide, DIPEA = diisopropylethylamine, Boc = *tert*-butyloxycarbonyl, HOBt = 1-hydroxybenzotriazole.

framework was installed by releasing **33** from the solid support and regioselective epimerization under basic conditions. The *cis* isomers formed underwent a Dieckmann cyclization to β-ketoesters **35**. The resulting macroline-inspired compound collection of about 100 molecules included potent inhibitors of the mycobacterial tyrosine phosphatase MptpB.

In a different natural product inspired synthesis, the diaza-bridged tetracyclic indole scaffold, which is part of many alkaloids (Scheme 4), was targeted.[48] The marine alkaloid yondelis (ET-743, Scheme 4), which contains the diaza-bridged scaffold, was granted orphan drug status in 2005 by the FDA for the treatment of ovarian cancer in the US. Although many natural products contain diaza-bridged systems, their natural scarcity and their complexity have limited their development as antitumor drugs. To access compound classes with diaza-bridged cyclic structural motifs that may display various biological activities (Scheme 4 b),[49] resin-bound tryptophan acetal **36** was deprotected and acylated to yield cyclization precursors **37**. The final regio- and diastereoselelctive cyclization was performed in neat formic acid, which led to the simultaneous release from the solid support and a Pictet–Spengler cyclization via in situ generated cyclic iminium ions. The diaza-bridged molecules **38** were obtained as single diastereomers in high yields and with high purities. The use of Fmoc-protected tryptophan and Fmoc-protected (*O*-diTBS)DOPA as substrates led to the preparation of a 384-member library of 3,9-diazabicyclo-[3.3.1]non-6-en-2-one skeletons, fused with indole and dihydroxybenzene, and diversified at two bridging nitrogen atoms.

Tricyclic benzopyrones, wherein a benzopyrone ring is fused to further heterocycles (**39** and **40**, Scheme 5), were found to be inhibitors of metallo-β-lactamases and thus potential antibacterials with activity against drug-resistant bacterial strains.[50] Inspired by these natural products and targeting the tricyclic benzopyrone core, a novel [4+2] annulation strategy was developed to generate a focused
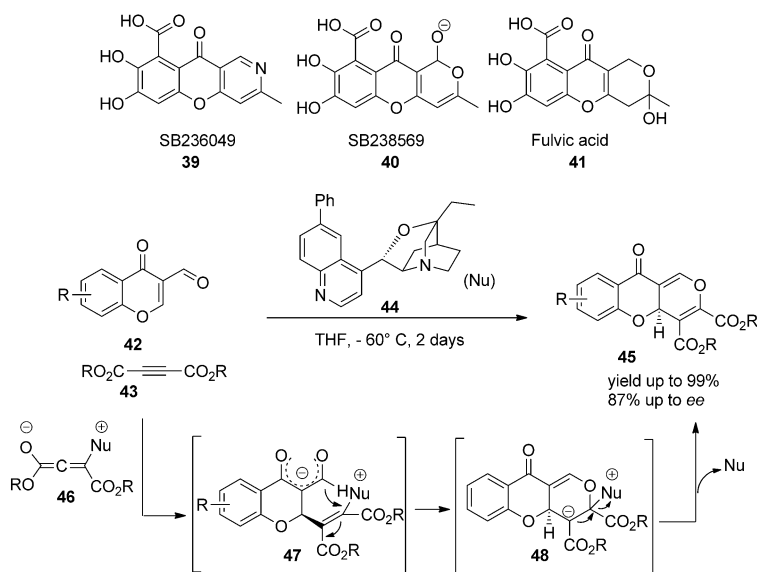
collection of tricyclic benzopyrones (**45**, Scheme 5).[51] The annulation between two electron-deficient systems, that is, oxadiene **42** and acetylenecarboxylates **43**, was facilitated by nucleophilic catalysis with tertiary phosphines or amines. Thus, the zwitterion **46** formed by the addition of organocatalyst **44** to alkynes **43** underwent a reaction sequence of Michael addition/Michael addition/elimination to generate the desired target structure. The use of cinchona-derived β-isocupreidines as catalysts provided an enantioselective route to (*S*)-**45**.

The synthesis of natural product inspired compound collections frequently requires multistep synthesis sequences to generate natural product like structural complexity. This demand often hinders the synthesis of medium-sized or large libraries and calls for the development of complexity-generating reactions that rapidly and efficiently generate complex molecular skeletons based on natural products.[52] Cascade or domino reaction sequences can provide efficient solutions to meet this challenge. For example, an efficient synthesis[53] of pyrroloisoquinolines related to the lamellarin alkaloids (Scheme 6), a family of marine natural products with a highly substituted pyrroloisoquinoline core and including inhibitors of human topoisomerase I and HIV-1 integrase,[54] made use of a domino synthesis. A silver(I)-catalyzed cycloisomerization of alkynyl *N*-benzylidene glycinates **49** to an azomethine ylide **52** followed by dipolar cycloaddition with the acetylenedicarboxylates gave rise to intermediates **53**. Isomerization followed by oxidative aromatization provided the pyrroloisoquinolines **51** in an efficient one-pot procedure.

Similarly, a cascade reaction sequence involving silver-catalyzed cycloisomerization of acetylenic aldehydes as the key transformation was recently explored in the synthesis of diverse alkaloid ring systems.[55] Indoloisoquinolines, a medicinally significant class of molecules known for their anticancer properties, were readily and efficiently generated by using this cascade approach. Thus, the imine generated from an acetylenic benzaldehyde **55** and an aniline with a pendant nucleophile **56** underwent the key silver-catalyzed cycloisomerization reaction under microwave conditions to yield the isoquinolinium cations **58**. A nucleophilic attack from the pendant nucleophile onto the iminium cation provided the intermediate **59**, which underwent decarboxylative aromatization to yield the target indolo[2,1-*a*]iso-quinolines **57** in good yields (Scheme 7 a).

The marine natural products homofascaplysin C and CDK-4 inhibitor fascaplysin were synthesized according to this method (Scheme 7 b). A microwave-assisted silver-catalyzed cascade cyclization of Boc-protected 3-ethynylindole-2-carbaldehyde (**60**) as a common precursor and aniline **61** yielded the pentacyclic core **62**. Formylation of **62** with POCl₃ cleanly provided homofascaplysin C, while oxidation of the pentacyclic core **62** with peracetic acid followed by treatment with acid efficiently yielded fascaplysin.
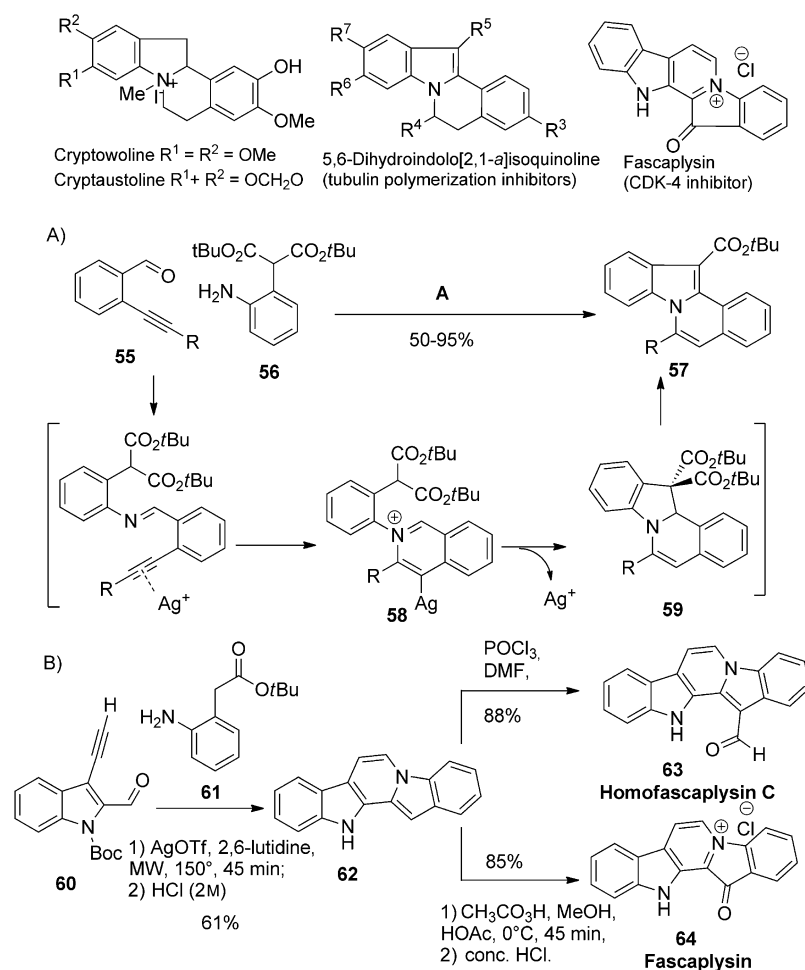
Selected scaffolds of additional compound collections inspired by natural products that pro-

Scheme 5. Synthesis of natural product inspired tricyclic benzopyrones.

Lamellarin D triacetate

Lamellarin α-20-sulfate
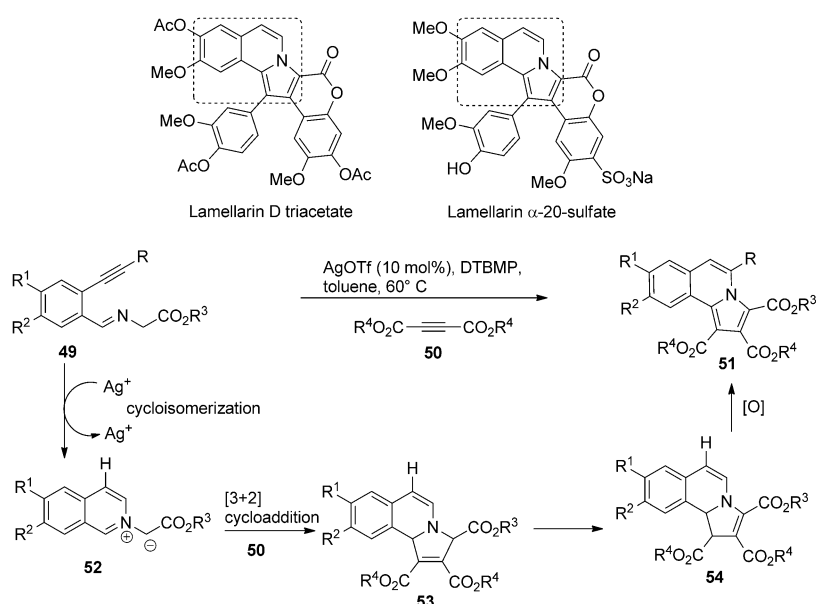
Scheme 6. Cascade synthesis of lamellarin-inspired molecules. DTBMP = 2,6-di(*tert*-butyl)-4-methylpyridine.

Cryptowoline $R^1 = R^2 = OMe$
Cryptaustoline $R^1 + R^2 = OCH_2O$

5,6-Dihydroindolo[2,1-*a*]isoquinoline
(tubulin polymerization inhibitors)

Fascaplysin
(CDK-4 inhibitor)

Scheme 7. Cascade synthesis of alkaloid-based compound collections. MW = microwave.

vided inhibitors or probes for biological applications are summarized in Figure 16. To obtain enantiopure natural product inspired α-β-unsaturated lactones, the hetero-Diels–Alder reaction of oxygen-substituted dienes with a glyoxylate in the presence of a chiral titanium catalyst yielded the desired dehydrolactones with high enantiomeric and diastereomeric ratios. Biological evaluation of these compounds in cell-based assays yielded new modulators of cell-cycle progression, and inhibitors of viral entry into cells were identified.[56] In another approach, the hetero-Diels–Alder reaction between a resin-bound aldehyde and a Danishefsky diene in the presence of chiral catalysts was employed to generate the lactones in high yield and high enantiomeric excess. The lactones were further modified on a solid phase to yield a natural product inspired compound collection based on the tetrahydropyran scaffold (**66**, Figure 16).[57]

Melophlin A and B are tetramic acid natural products that reverse the morphology of HRas-transformed NIH3T3 fibroblasts at a concentration of 5 µg mL^{-1}.[58] A melophlin-inspired compound collection was generated to identify their biological target and their role in the Ras signaling network (**67**, Figure 16). Biological evaluation and subsequent chemical proteomics investigations revealed that melophlin A unexpectedly targets dynamins in cells, and thereby modulates signal transduction through the Ras network indirectly by preventing endocytosis of MEK, a downstream target of Ras signaling.[59]

β-Lactones occur in various natural products and were used as scaffolds for the synthesis of palmostatins (**68**, Figure 16). Palmostatin B was developed as an inhibitor of acyl protein thioesterase 1 (APT1),[33] and was successfully employed to establish the role of this thioesterase in regulating the localization, intracellular transport, and signaling of the *S*-palmitoylated H- and N-Ras proteins in general (see Section 3.2).[60]

Cyclopeptide core structures are frequently found in natural products. The brunsvicamides are modified cyclopeptides from cyanobacteria, cyclized through the ε-amino group of a D-lysine unit and functionalized with urea groups. They show potent carboxypeptidase inhibitory activy. A collection of modified brunsvicamides was synthesized by varying the amino acid residues and stereochemistry pattern in a combined solution- and solid-phase approach.[61] The small library was biochemically evaluated for inhibition of carboxypeptidase A. The results revealed
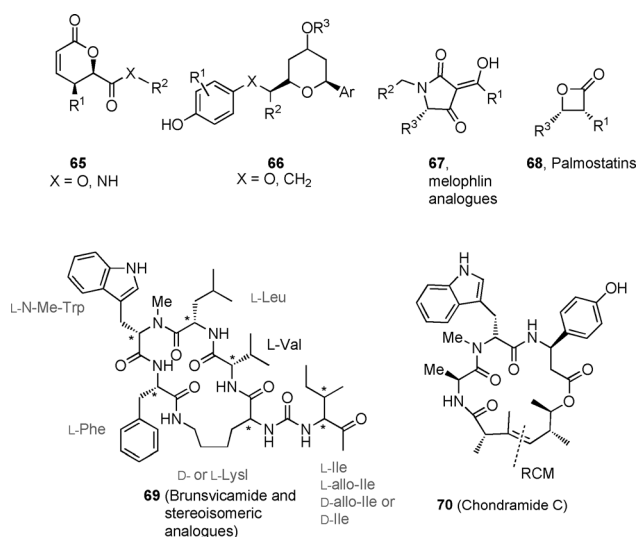
Figure 16. Compound collections based on the BIOS approach.

the significance of different amino acid residues and especially the high relevance of the lysine stereochemistry for inhibitory activity. Furthermore, a synthesis of chondramide C inspired cyclopeptides was developed and applied to build up a library of potential modulators of actin filaments.[62] The key macrocylization step was realized through ruthenium-catalyzed ring-closing metathesis (RCM), which in the course of the synthesis of a library produced discernible trends in metathesis reactivity and *E/Z* selectivity. The inhibitory effects of the synthesized compounds on growth were quantified and structure–activity correlations established, which appear to be in good alignment with relevant biological data from natural products. Thus, a number of potent non-natural and simplified analogues were identified for further in-depth studies of the mode of action, especially into the relationship between the cytotoxicity of these compounds and their actin-perturbing properties.

In addition to these illustrating examples, various studies have been reported that describe the successful synthesis of natural product inspired compound collections (for a comprehensive review, see Ref. [38,39] for a review of selected examples, see Ref. [9ab]). It can be safely concluded on the basis of this collective effort from the scientific community that the available synthesis methods allow for the reliable and speedy synthesis of natural product inspired compound collections.

3.2. Application of BIOS in Inhibitor and Ligand Development and Chemical Biology

3.2.1. Application of PSSC in the Development of an APT1 Inhibitor

The H- and N-Ras proteins are *S*-palmitoylated membrane-bound GTPases critically involved in growth-factor signaling across the plasma membrane. Mutations in Ras proteins are found in approximately 30 % of all cancers. The reversible removal and attachment of palmitic acid from and

to cysteine residues at the C terminus (depalmitoylation and palmitoylation) of the H- and N-Ras isoforms control their membrane attachment and specific localization. The dynamics of palmitate turnover is crucial to H- and N-Ras signaling as well as to establish a cycle of Ras-trafficking between the plasma membrane and the Golgi (Figure 17).[63] Acyl protein thiosterase 1 (APT1) was the only enzyme known to depalmitoylate H- and N-Ras. However, its role in the Ras cycle was unclear. Since shuttling of Ras between its different cellular locations occurs on the second-to-minute time scale, a chemical–genetic approach making use of rapid APT1 inhibition appeared to be particularly suitable to unravel the role of APT1 in the dynamic Ras cycle. Since an inhibitor suitable for this purpose was not available, PSSC was employed for the development of an APT1 inhibitor.
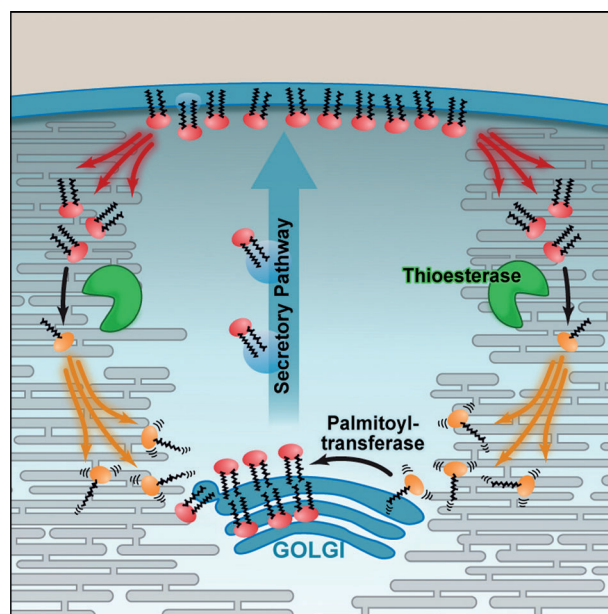

Figure 17. The dynamic nature of the Ras cycle. Reproduced from *Cell* **2010**, *141*, 458–471.

A search for subfold similarity based on the ligand binding site of APT1 by PSSC analysis yielded dog gastric lipase as a hit, with a high structural similarity despite its relatively low sequence similarity (below 25 %).[33] Analysis of an overlay of both active-site structures showed a very similar spatial arrangement of the catalytic residues (Figure 18). This finding suggested that compounds similar to lipase inhibitors might be APT1 inhibitors. The natural product derived marketed lipase inhibitor tetrahydrolipstatin (Orlistat) contains a β-lactone, which is attacked by and inhibits the enzyme by formation of an acyl enzyme intermediate. On the basis of this analysis, a collection of β-lactones was synthesized and the most potent compound termed palmostatin B was analyzed in detail (Figure 18). Palmostatin B competitively inhibits APT1 with IC$_{50}$ = 670 nM through reversible acylation of the nucleophilic serine in the catalytic triad of the enzyme. The resulting palmostatin B/APT1 complex hydrolyzes slowly, and the compound itself has a half-life of 58 h in aqueous
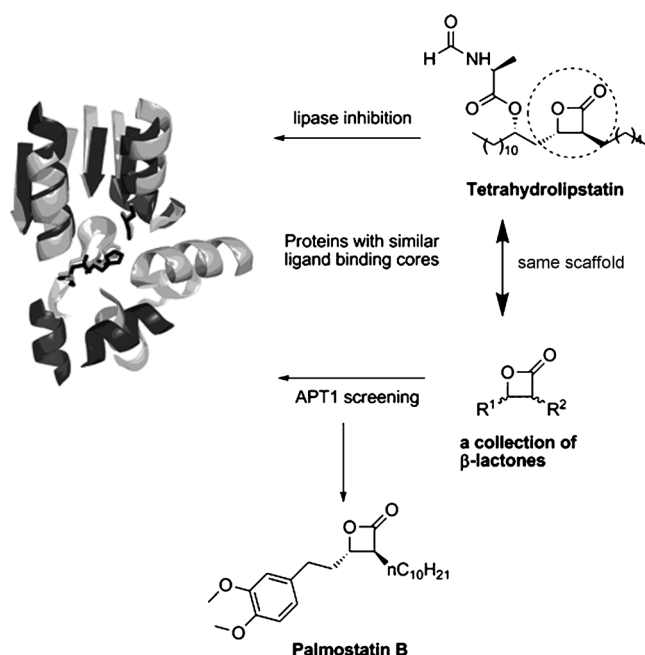
Figure 18. Protein structure similarity cluster of APT1 (dark gray) and gastric lipase (light gray) and the logic for the synthesis and screening of a β-lactone collection that yielded the APT1 inhibitor palmostatin B.

solution. Palmostatin B is sufficiently soluble and cell-permeable to make it a useful tool for the study of APT1 function in the Ras acylation/deacylation cycle.

Accordingly, palmostatin B was employed in a series of biochemical and live-cell investigations, including time-resolved fluorescence microscopy studies, which proved that the compound interacts with APT1 in cells, is selective for APT1 over other intracellular hydrolases, and inhibits the depalmitoylation of H- and N-Ras in cells. Palmostatin B perturbs the cellular acylation cycle at the level of depalmitoylation and thereby leads to loss of precise, steady-state membrane localization of the palmitoylated Ras proteins through entropy-driven distribution of the proteins among cellular membranes (Figure 17). In this way, it counterintuitively attenuates H-/N-Ras signaling and induces partial phenotypic reversion of H-Ras-transformed MDCK-F3 cells to the nontransformed phenotype.

The study clearly identified APT1 as a decisive thioesterase in the acylation cycle and suggests that APT1 may be a novel anticancer target. Palmostatin B may be a valuable starting point for the development of modulators of pathological signaling by palmitoylated Ras proteins.

3.2.2. Application of SCONP in the Identification of Novel Phosphatase Inhibitors

Protein phosphatases are key regulators of innumerable biological processes and targets in drug discovery programs, for example, in diabetes and anticancer research.[64] However, the inhibition of phosphatase in cells and in vivo has proven to be difficult and, therefore, novel classes of phosphatase inhibitors are in high demand.

In an attempt to identify such novel compounds, a collection of 354 natural products was screened for their inhibition of several phosphatases.[65] Surprisingly, the pentacyclic indole alkaloid yohimbine was identified as an inhibitor ($IC_{50} = 22.3$ μM) of the dual-specificity phosphatase Cdc25A, which has been considered as an anticancer target. Since the synthesis of a compound collection based on the pentacyclic yohimbine scaffold would be a major challenge, the natural product was subjected to brachiation and SCONP analysis. This analysis led to tetra-, tri-, and bicyclic natural product scaffolds, which inspired the synthesis of a compound library (Figures 4 and 19).
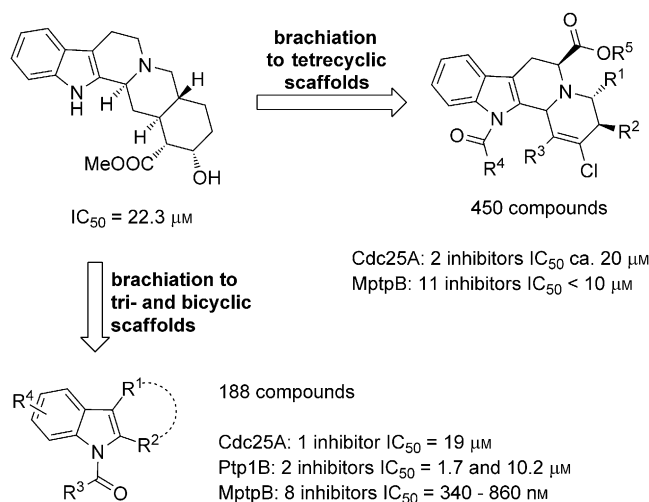

Figure 19. Brachiation through the indole branch of the natural product scaffold tree and development of novel natural product inspired classes of phosphatases inhibitors.

A collection of 450 tetracyclic indoloquinolizidines was synthesized as shown in Scheme 4, and additionally a collection of 188 tri- and bicyclic indole derivatives was synthesized by means of a Fischer indole synthesis and a resin-capture-and-release strategy. Biochemical analysis of the compound collection for inhibition of Cdc25a revealed two tetracyclic and one tricyclic compound displaying IC_{50} values comparable to that of the natural product itself. Subsequent screening for the inhibition of further phosphatases identified novel inhibitors of protein tyrosine phosphatase 1B (PTP1B), a major target in diabetes research, as well as nanomolar inhibitors of the mycobacterial protein tyrosine phosphatase B (MptpB), which is a promising target for the discovery of novel antituberculosis drugs.

These results demonstrate successful brachiation through the N-heterocyclic indole branch of the SCONP tree. They show that the BIOS approach allows substantial reduction of the molecular complexity with retained bioactivity, and that BIOS offers the opportunity to discover novel readily accessible inhibitor classes based on complex structures of natural products.

3.2.3. *Combining PSSC and SCONP: Decalins as Selective 11βHSD1 Inhibitors*

High levels of glucocorticoids, which are steroid hormones that regulate glucose metabolism, may cause the development of the metabolic syndrome.[66] The active glucocorticoid cortisol is produced by the 11β-hydroxysteroid dehydrogenase 1 (11βHSD1) catalyzed reduction of cortisone. 11βHSD1 is mainly expressed in the liver, adipose tissue, and brain. In the kidneys, 11βHSD1 catalyzes the inactivation of cortisol by oxidation to cortisone, thereby protecting the body from cortisol-induced hypertension. In mice, global genetic ablation of 11βHSD1 leads to increased insulin sensitivity and resistance to diet-induced obesity, hyperglycemia, and dislipidemia. These results suggest that selective 11βHSD1 inhibitors may be useful in the treatment of type 2 diabetes and metabolic syndrome as well as the prevention of atherosclerosis.[67]

Efforts by pharmaceutical and biotechnology companies have led to several nonsteroidal 11βHSD1 inhibitors with beneficial effects in animal models of atherosclerosis and type 2 diabetes, and the search for isoenzyme-selective 11βHSD inhibitors is ongoing. The synergistic combination of PSSC and SCONP has resulted in new types of selective 11βHSD1 inhibitors with cellular activity.[14]

Figure 20. A) SCONP analysis of glycyrrhetinic acid and dysidiolide and rationale for the identification of the selective 11βHSD1 inhibitor **71**. B) Superimposed catalytic sites of Cdc25A (red), 11βHSD1 (green), and AChE (blue). The key catalytic residues, Cys-430 (Cdc25A), Tyr-183 (11βHSD1), and Ser-200 (AChE) are shown in space-filling representation.

By using PSSC, 11βHSD1 and 11βHSD2 were assigned to a cluster that also contains the dual specificity phosphatase Cdc25 A and acetylcholine esterase (AChE, Figure 13). Although Cdc25A and AChE are mechanistically unrelated to the 11βHSDs and the sequence identity is low (< 10%), the subfolds of their catalytic sites and positions of their catalytic residues show very good overlap (Figure 20 B).

At the time of the analysis, the structure of 11βHSD1 had not been determined and a homology model was used to generate the PSSC. Later the structure of the enzyme became available.[68] Comparison confirmed the validity of the homology model and demonstrated that—in principle—high-resolution crystal structures of proteins may not necessarily be required for an initial hypothesis-generating PSSC analysis.

Subsequently, the natural Cdc25A inhibitor dysidiolide and the 11βHSD ligand glycyrrhetinic acid (GA) were analyzed by using the SCONP tree. Stepwise deconstruction of the pentacyclic scaffold of GA led to the bicyclic 3,4-dehydrodecalin scaffold **IV**, whereas SCONP analysis of dysidiolide resulted in bicyclic parent scaffold 1,2-dehydrodecalin **VI** (Figure 20 A). Since **VI** can be considered an alternative subscaffold of GA, a natural product inspired library of 483 decalins based on scaffold **VI** was synthe-
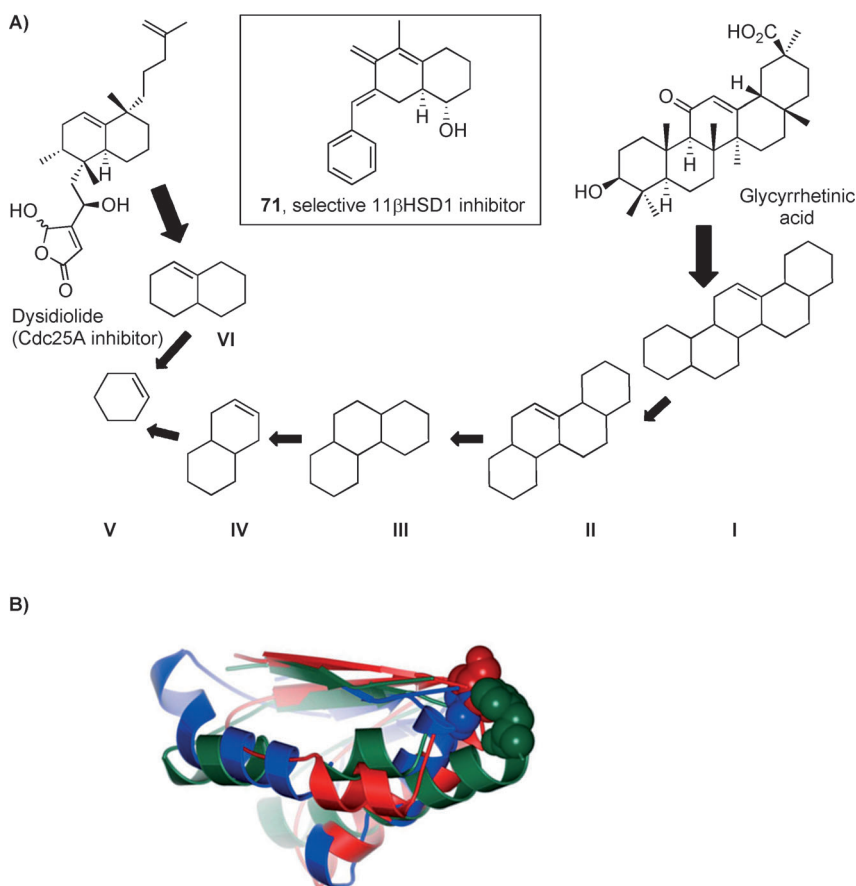
sized.[69] In addition to several low micromolar AChE inhibitors, this library contained three highly isoenzyme-selective, nanomolar 11βHSD1 inhibitors. The selective 11βHSD1 inhibitor **71** was subsequently shown to inhibit cortisol-mediated glucocorticoid receptor translocation of HEK-293 cells to the nucleus at low micromolar concentrations, thus indicating that this new compound also inhibits 11βHSD1 in cells.

4. Summary and Outlook: Where Do We Come from and Where Are We Going?

Bioactive small molecules offer unique and often unprecedented opportunities for the analysis of complex biological phenomena by rapidly, temporarily, conditionally, selectively, and tunably perturbing but not changing biological systems. Rather than targeting the whole chemical space, the key to the discovery of bioactive small molecules is the development and application of methods that allow one to identify, chart, and navigate biologically relevant chemical space. Ultimately, such methods must enable prospective exploration of chemical space and prediction of bioactivity for particular compound classes.

To approach this goal we have introduced a Structural Classification of Natural Products (SCONP). The underlying frameworks of natural products provide evolutionary-selected chemical structures that encode the properties required for binding to proteins, and their structural scaffolds represent the biologically relevant and prevalidated fractions of chemical space explored by nature in evolution. Consequently, it is to be expected that compound collections designed on the basis of the structures of natural products will be enriched in biochemical and biological activity. The treelike hierarchical arrangement of natural product scaffolds in SCONP provides an idea- and hypothesis-generating tool for the design and synthesis of compound collections.

Furthermore, we have introduced Protein Structure Similarity Clustering (PSSC) as an analogous hypothesis-generating method that employs the conservation of protein structure in evolution and structural similarity among protein binding sites to identify new ligand types for proteins of interest.

Both SCONP and PSSC suggest that nature synergistically employed elements of conservatism and of diversity in the evolution of the small molecules it made and the proteins employed to make them and to which they bind when they fulfil their biological function. At the level of the scaffolds, nature was conservative in both the small-molecule and the protein world. In both cases this element was complemented by a level of diversity represented by the substituents of small molecules and their attachment sites as well as the side chains of the amino acids in the ligand-sensing protein cores. The matching of scaffold architecture and of substituent structure as well as positioning will enable the design and identification of biologically relevant small-molecule classes.

Both SCONP and PSSC inspire the selection of compound library scaffolds on the basis of the relevance to and prevalidation by nature. We refer to synthesis efforts based on these criteria as Biology-Oriented Synthesis (BIOS). In BIOS, either a SCONP or PSSC analysis may be employed separately or synergistically for the generation of hypotheses and ideas and to guide the synthesis of compound collections. In BIOS, focused diversity around a biologically relevant starting point in vast structure space is generated. BIOS may build on the diversity created by nature in evolution and aim at its local extension in areas of proven relevance by means of natural product inspired or derived compound collections. However, non-natural product scaffold types with proven biological relevance are also fully valid starting points for BIOS approaches, that is, BIOS is not restricted to natural product scaffold classes. It calls for biological relevance as the guiding argument rather than occurrence in nature. Thus, BIOS may yield new opportunities for the discovery of unprecedented protein ligand and inhibitor classes with relatively high hit rates in comparably small compound libraries. Through brachiation along the branches of scaffold trees, it may also serve as a hypothesis generator to arrive at structurally simpler scaffolds that retain the same kind of bioactivity, often with graded potency and selectivity.

BIOS provides a mainly structure-based, chemocentric view to the problem of identifying bioactive small molecules and chemotypes. The basic idea for the development of SCONP, PSSC, and BIOS was born and shaped in the second half of the 1990s, namely, at a time when the initial wave of combinatorial chemistry and high-throughput screening had swept through industry and academia, when very large compound libraries had been synthesized mainly based on criteria of chemical feasibility and commercial availability of building blocks. At that time, the picture had begun to emerge that high-volume screening of such libraries resulted in very low hit rates compared to the approximately two orders of magnitude higher hit rates from historic compound collections in the pharmaceutical industry and from pure collections of natural products.[9c] However, even with full recognition of this discrepancy—for which there was no straightforward explanation at hand—most pharmaceutical companies had progressed to eliminate natural products from their screening libraries. Natural products appeared to be structurally too complex to pursue and synthesize, too large, and often not available in sufficient amounts from natural sources for further development. The mostly technology-driven development of high-throughput techniques seemed suitable to meet the need for an increasing number of hits, leads, drugs, and also chemical probes for biological investigations.

However, it rapidly became clear that this could not be achieved by simply increasing the number of screens, libraries, and data points, but rather that high-quality chemical libraries were needed that met additional criteria, such as biological relevance, drug-likeness, structural complexity, and diversity. Aware of these facts and developments, in particular the excellent performance of natural products and the contradictory simultaneous decision to eliminate them from drug discovery in industry, we began to ask whether there might be a logic and method to reduce the structural complexity of natural products but retain their bioactivity. Could an underlying logic be developed to systematically analyze the structural complexity of natural products, their relationship to each other, and also to the structural diversity in the binding sites of target proteins? Could such a logic be used to inspire the synthesis of compound libraries and would it be chemically feasible to synthesize compound collections with structures approaching the complexity of natural products in the required formats, such as solid-phase synthesis? And if so, would these libraries also approach the performance of natural products in biochemical and biological screens, namely, would they meet quality criteria and deliver relatively high hit rates at comparably small library size, thereby reducing the need for engagement in high-throughput techniques?

If successful, such a logic and approach could inspire and promote the reintroduction of natural product structures into the discovery and development of candidate molecules in both medicinal chemistry and chemical biological research—however, then with a firm grip on the molecular complexity and progressable synthesis routes.

In response to these and related questions SCONP, PSSC, and BIOS were developed as the guiding underlying logic to identify, analyze, and hierarchically arrange biologically relevant scaffold classes to inspire synthesis efforts and to even prospectively assign the kind of bioactivity for compound classes. The results gleaned from BIOS libraries

indicate that reduction of the structural complexity of natural products and also non-natural products with a retained type of bioactivity is indeed possible and that this is valid for all current major classes of drug targets.

Brachiation following the logic of BIOS differs from attempts to simplify natural product structures on the basis of chemistry arguments alone, for example, higher synthesis efficiency or retrosynthetic consideration. In BIOS, brachiation needs to follow lines of biological relevance defined, for example, by the occurrence of smaller scaffolds in nature or available bioactivity data. The chemistry required to synthesize compound collections with smaller scaffolds then has to be selected accordingly. Thus, in BIOS, the selection of the synthesis targets follows biological arguments and selection criteria.

BIOS-based libraries are small and focused, and show relatively high hit rates. Our own contributions to the synthesis of natural product inspired and derived compound collections and a variety of excellent results reported by various research groups worldwide[9a, 38, 39] allow us to conclude safely that organic synthesis methods are sufficiently developed and powerful to grant reliable and flexible access to such compound collections with reasonable effort both in academic and industrial settings.

The chemical effort required to synthesize such libraries may be high and require more time for development, but it will result in compound collections endowed with biological relevance. In a sense, BIOS calls for a more-intense investment in the chemistry part of the development of bioactive small molecules, which will pay off because it will yield better molecules for biology research.

It should be noted, however, that BIOS also reverses the reasoning and inspiration for the establishment of synthesis projects driven, for example, by the desire to develop novel methods or to achieve a total synthesis of a given natural product. The bioactivity and relevance of particular natural product scaffolds determine the synthesis target, and chemical synthesis strategies and methods will have to be adapted to meet the resulting requirements.

In purely synthetic investigations it is often the method that dictates which natural product will be synthesized. Total syntheses require that a given natural product has to be made in all details, while BIOS reduces natural product structure to the scaffold, its equipment with different substituents, and variation of stereochem-

istry. Precise synthesis of the guiding natural products in all details is appreciated but not required.

As mentioned above, at the time when the ideas leading to the development of SCONP, PSSC, and BIOS had begun to take shape, the pharmaceutical industry was in the process of eliminating natural products from their screening collections and to discontinue and spin-out natural product research departments. With a few exceptions, substantial collections of natural products are today predominantly in the hands of smaller, specialized companies such as InterMed Discovery GmbH and AnalytiCon GmbH. Natural products were considered too big, complex, and synthetically nontractable to fit into the discovery and development time lines and pipe lines of pharma companies. However, opinion may frequently have dominated over facts in this reasoning and the subsequent processes. Thus, statistical analysis of the SCONP tree showed that more than half of all natural products have scaffolds with two, three, or four rings, and that their van der Waals volumes match the lower end of the sizes of cavities found in and on proteins. Consequently the majority of natural products have just the right size (!) to serve as starting points for hit and lead discovery as well as for development programs including the attachment of further substituents (Figure 21). Also, the SCONP analysis can, in principle, reveal the attachment sites of substituents on the scaffolds, thus further inspiring design.

Beyond this, the successive deconstruction of natural product scaffolds can be carried out such that the smaller scaffolds are "fragment-like" in the sense of fragment-based drug discovery.[71] This analysis, in a sense, reveals the "fragments of nature" and will further fuel the synthesis of natural product inspired compound collections and design.

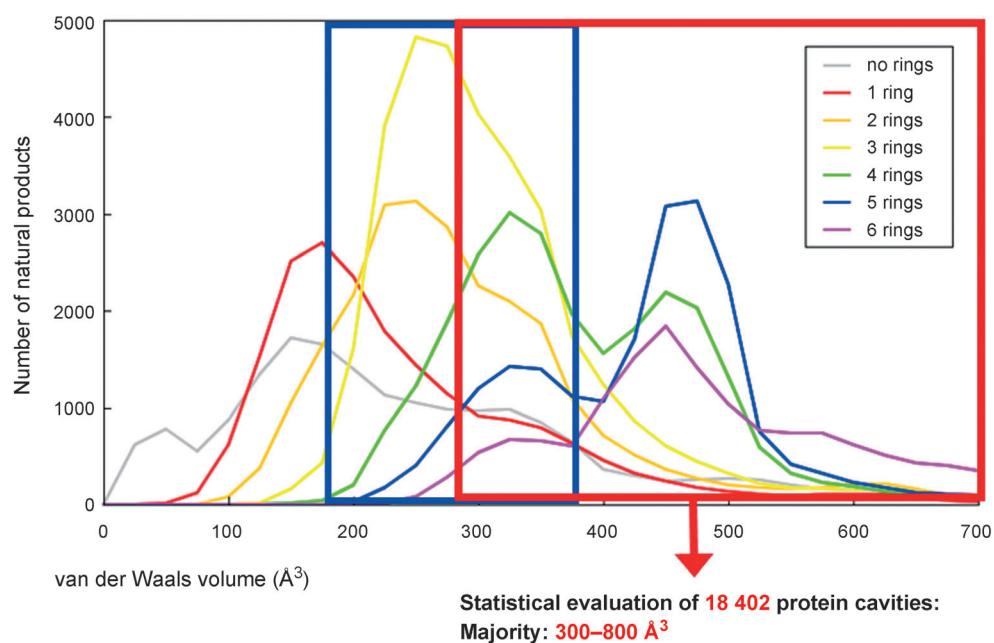**Statistical evaluation of 18 402 protein cavities:
Majority: 300–800 Å³**

Figure 21. Comparison of the van der Waals volume of natural product scaffolds containing different ring numbers with the volumes identified in proteins.[70] The volume of natural product scaffolds with 2–4 rings are at the lower edge of the sizes of cavities in proteins, thus suggesting that these scaffolds are not too big for further compound development.

Finally, the efforts to synthesize natural product inspired compound collections, as summarized above, have shown that compound libraries approaching the complexity of natural products are indeed within reach and that currently available synthesis methods are sufficiently powerful to reach the goal of making them available in industrial and academic formats.

Taken together, these results and the conclusions emanating therefrom suggest that it may be prudent, indicated, and wise for the pharmaceutical industry to reintegrate library design based on natural products, synthesis, and screening into their research and development programs. This need not be in the former way of focusing on the individual natural products themselves. Natural product inspired and derived compound collections and natural product derived fragment-based design should meet the needs and restrictions that often have to be accepted in an industrial environment. The logic of BIOS also shows that the use of natural products alone as inspiration to identify, chart, and navigate biologically relevant chemical space is not sufficient and leaves "holes" in chemical space. Instead, it is necessary to expand the analysis to as many bioactive and, therefore, biologically relevant compound classes as possible—be they natural products or not—ideally to all known bioactive compounds. This need is convincingly highlighted by one of the most successful examples of the structural simplification of a natural product to smaller scaffolds that retain the same kind of bioactivity. Morphine cannot be placed into the SCONP tree because a suitable tetracyclic scaffold has not been identified in nature. However, non-natural tetracyclic morphine-derivatives were actually developed as marketed drugs (Figure 22).

BIOS was initially developed using natural products as guiding prevalidated examples, but is not restricted to them. Natural products reflect the solution to identify biologically relevant areas in chemical space developed in evolution. However, it is clear that there are and will be other solutions, not explored by nature.

For this expansion of the coverage of biologically relevant chemical space it will be necessary to gain access to substantially larger data sets that correlate structure with bioactivity than is assembled in the WOMBAT database used in our analysis. Very recently the publicly available CHEMBL database was launched on the internet[72] which covers a wealth of bioactivity data reported in the scientific literature. In addition, PubChem also provides a large data set that is accessible for analysis. Coverage of these databases in addition to DNP and WOMBAT should allow a substantially advanced analysis of biologically relevant chemical space.

A further step in the development of such resources may consist of the application of automated full text mining of the entire scientific literature, including correlation of the chemical structure and bioactivity of small molecules. The largest sources of high-quality data, however, are only available inside the major pharmaceutical companies, who over the decades have investigated millions of compounds in hundreds of biochemical and biological screens. If access to these databases could be gained and if they could be subjected to analyses in the sense of the BIOS approach, it is to be expected that numerous novel research projects would be inspired that would fuel chemical biology and medicinal chemistry research programs, and potentially lead to the faster discovery and development of novel and better drugs.

We do not expect this to happen within the near future. Instead, academic research will be inspired by analysis of databases such as CHEMBL and PubChem. However, we also wonder whether the pharmaceutical companies know where the holes are in their compound collections, databases, and patents?
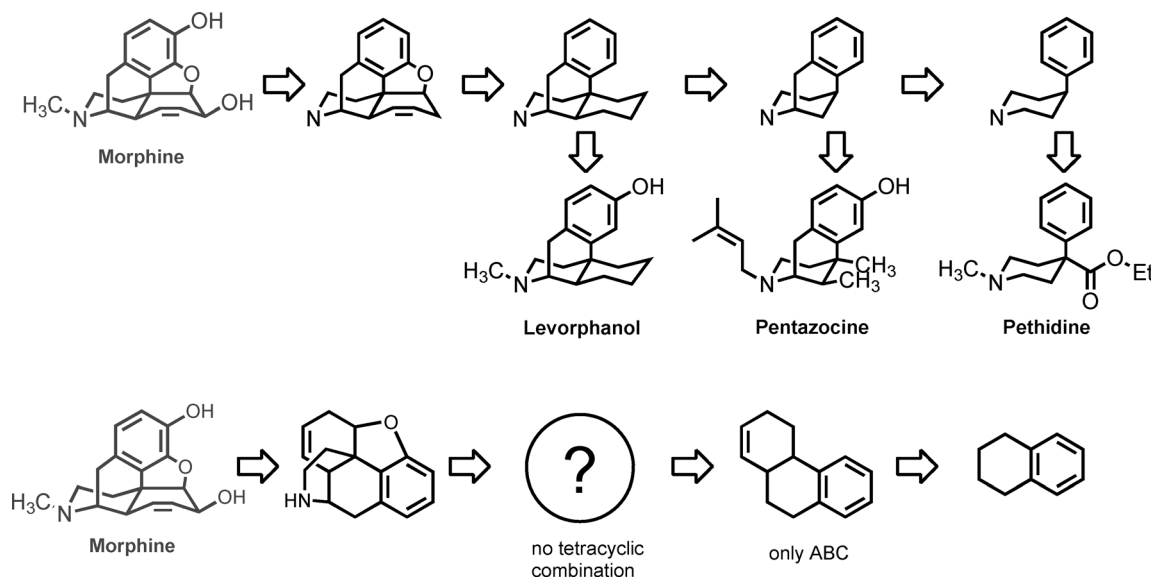

Figure 22. The dilemma encountered upon attempting to place morphine and its relatives with smaller scaffolds into the natural product tree. Morphine has been deconstructed in pharmaceutical development to give marketed drugs with tetracyclic and tricyclic scaffolds. However, analysis of morphine in SCONP reveals that no tetracyclic scaffold derived from morphine occurs in nature and that only one tricyclic scaffold is known. Thus, a "hole" in natural product chemical space is not shared by a "hole" in natural product bioactivity space. This dilemma suggests that the SCONP analysis must be complemented by the inclusion of bioactive non-natural products for a better analysis of biologically relevant chemical space.

5. Summary

Bioactive small molecules offer unique opportunities to acutely perturb and analyze complex biological systems. Their discovery calls for the development of methods that allow one to identify, chart, navigate, and populate biologically relevant chemical space. Biology-oriented synthesis (BIOS) approaches this problem by means of a chemocentric analysis of the structures of the ligand-sensing cores embedded in protein domain folds and the scaffold structures of natural product classes generated through evolution as well as further non-natural bioactive compound classes. Protein Structure Similarity Clustering (PSSC) and a Structural Classification of Natural Products (SCONP) and its extension to bioactive non-natural products were developed for this analysis. Either applied alone or synergistically, these bio- and cheminformatic methods serve as hypothesis-generating tools to identify small-molecule scaffold classes endowed with biological relevance. Such scaffolds fuel synthesis programs to generate small or medium-sized compound collections, for example, inspired by natural product structures, with focused diversity around a biologically relevant starting point in vast chemical structure space. The analysis of biologically relevant chemical space is facilitated by the Scaffold Hunter, an intuitively accessible and interactive software that arranges scaffolds hierarchically according to chemical structure, and by a method for bioactivity-guided navigation of chemical space.

Natural product inspired compound collections with focused chemical diversity can be synthesized efficiently by means of multistep solution and solid-phase methods, domino- and cascade reactions, as well as multicomponent reactions which are further facilitated by the use of polymer-immobilized scavenging reagents and novel separation techniques. The natural product inspired compound collections synthesized according to the logic of BIOS prove to be enriched in bioactivity and yield inhibitors and modulators of bioactivity in biochemical and cell-based assays typically in the 0.2–1.5% range. They have been used successfully to analyze complex biological processes.

The successful development of the BIOS approach paves the way to employ the biological prevalidation of natural product structure by evolution in chemical biology and medicinal chemistry research, thereby overcoming limitations of synthetic tractability or accessibility of natural products and suggests that natural products, and compound collections inspired by them should be reconsidered in future drug discovery efforts.

[1] E. Zamir, P. I. H. Bastiaens, *Nat. Chem. Biol.* **2008**, *4*, 643–647.

[2] a) S. V. Frye, *Nat. Chem. Biol.* **2010**, *6*, 159–161; b) S. Wetzel, A. Schuffenhauer, S. Roggo, P. Ertl, H. Waldmann, *Chimia* **2007**, *61*, 355–360; c) K. Grabowski, K. H. Baringhaus, G. Schneider, *Nat. Prod. Rep.* **2008**, *25*, 892–904.

[3] http://www.nature.com/nchembio/chemical_probes.html.

[4] S. L. Schreiber, *Nat. Chem. Biol.* **2005**, *1*, 64–66.

[5] a) C. M. Dobson, *Nature* **2004**, *432*, 824–828; b) R. S. Bohacek, C. McMartin, W. C. Guida, *Med. Res. Rev.* **1996**, *16*, 3–50; c) R. S. Bon, H. Waldmann, *Acc. Chem. Res.* **2010**, *43*, 1103–1114.

[6] a) G. Klebe, *Drug Discovery Today* **2006**, *11*, 580–594; b) K. J. Simmons, I. Chopra, C. W. G. Fishwick, *Nat. Rev. Microbiol.* **2010**, *8*, 501–510.

[7] a) M. Rupp, T. Schroeter, R. Steri, H. Zettl, E. Proschak, K. Hansen, O. Rau, O. Schwarz, L. Muller-Kuhrt, M. Schubert-Zsilavecz, K. R. Muller, G. Schneider, *ChemMedChem* **2010**, *5*, 191–194; b) P. Willett, *Drug Discovery Today* **2006**, *11*, 1046–1053; c) J. Rosén, J. Gottfries, S. Muresan, A. Backlund, T. I. Oprea, *J. Med. Chem.* **2009**, *52*, 1953–1962.

[8] D. J. Newman, G. M. Cragg, *J. Nat. Prod.* **2007**, *70*, 461–477.

[9] a) K. Kumar, H. Waldmann, *Angew. Chem.* **2009**, *121*, 3272–3290; *Angew. Chem. Int. Ed.* **2009**, *48*, 3224–3242; b) M. Kaiser, S. Wetzel, K. Kumar, H. Waldmann, *Cell. Mol. Life Sci.* **2008**, *65*, 1186–1201; c) R. Breinbauer, I. R. Vetter, H. Waldmann, *Angew. Chem.* **2002**, *114*, 3002–3015; *Angew. Chem. Int. Ed.* **2002**, *41*, 2878–2890; d) S. C. K. Sukuru, J. L. Jenkins, R. E. J. Beckwith, J. Scheiber, A. Bender, D. Mikhailov, J. W. Davies, M. Glick, *J. Biomol. Screening* **2009**, *14*, 690–699; e) D. H. Drewry, R. Macarron, *Curr. Opin. Chem. Biol.* **2010**, *14*, 289–298.

[10] F. Lovering, J. Bikker, C. Humblet, *J. Med. Chem.* **2009**, *52*, 6752–6756.

[11] B. E. Evans, K. E. Rittle, M. G. Bock, R. M. Dipardo, R. M. Freidinger, W. L. Whitter, G. F. Lundell, D. F. Veber, P. S. Anderson, R. S. L. Chang, V. J. Lotti, D. J. Cerino, T. B. Chen, P. J. Kling, K. A. Kunkel, J. P. Springer, J. Hirshfield, *J. Med. Chem.* **1988**, *31*, 2235–2246.

[12] R. I. Sadreyev, N. V. Grishin, *BMC Struct. Biol.* **2006**, *6*, 6.

[13] a) T. Henkel, R. M. Brunne, H. Muller, F. Reichel, *Angew. Chem.* **1999**, *111*, 688–691; *Angew. Chem. Int. Ed.* **1999**, *38*, 643–647; b) M. L. Lee, G. Schneider, *J. Comb. Chem.* **2001**, *3*, 284–289; c) M. Feher, J. M. Schmidt, *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 218–227.

[14] M. A. Koch, A. Schuffenhauer, M. Scheck, S. Wetzel, M. Casaulta, A. Odermatt, P. Ertl, H. Waldmann, *Proc. Natl. Acad. Sci. USA* **2005**, *102*, 17272–17277.

[15] *Dictionary of Natural Products*, Chapman & Hall/CRC Informa, London, **2005**.

[16] A. Schuffenhauer, P. Ertl, S. Roggo, S. Wetzel, M. A. Koch, H. Waldmann, *J. Chem. Inf. Comput. Sci.* **2007**, *47*, 47–58.

[17] K. O. Elliston, Talk at the 8th [BC]2 Basel Computational Biology Conference, **2010**.

[18] S. Renner, W. A. L. van Otterlo, M. D. Seoane, S. Mocklinghoff, B. Hofmann, S. Wetzel, A. Schuffenhauer, P. Ertl, T. I. Oprea, D. Steinhilber, L. Brunsveld, D. Rauh, H. Waldmann, *Nat. Chem. Biol.* **2009**, *5*, 585–592.

[19] S. Wetzel, K. Klein, S. Renner, D. Rauh, T. I. Oprea, P. Mutzel, H. Waldmann, *Nat. Chem. Biol.* **2009**, *5*, 696–696.

[20] *CTfile formats*, Symyx Technologies, San Ramon, CA, USA, **2007**; http://www.symyx.com/downloads/public/ctfile/ctfile.pdf.

[21] CambridgeSoft Corporation, **2006**.

[22] ISIS Draw, Symyx Technologies, San Ramon, CA, USA.

[23] a) M. Olah, M. Mracec, L. Ostopovici, R. Rad, A. Bora, N. Hadaruga, I. Olah, M. Banda, Z. Simon, M. Mracec, T. I. Oprea, *WOMBAT: world of molecular bioactivity, Vol. 23*, Wiley-VCH, Weinheim, **2005**; b) M. Olah, R. Rad, L. Ostopovici, A. Bora, N. Hadaruga, D. Hadaruga, R. Moldovan, A. Fulias, M. Mracec, T. I. Oprea, *Chemical Biology: From Small Molecules to Systems Biology and Drug Design, Vol. 1–3* (Eds.: S. L. Schreiber, T. M. Kapoor, G. Wess), Wiley-VCH, Weinheim, **2007**, pp. 760–786.

[24] a) H. Xu, S. M. Kahn, J. R. Peterson, E. Behar, F. B. S. Paerels, R. F. Mushotzky, J. G. Jernigan, A. C. Brinkman, K. Makishima, *Astrophys. J.* **2002**, *579*, 600–606; b) J. Inglese, D. S. Auld, A. Jadhav, R. L. Johnson, A. Simeonov, A. Yasgar, W. Zheng, C. P. Austin, *Proc. Natl. Acad. Sci. USA* **2006**, *103*, 11473–11478.

[25] a) W. Schulz, *Chem. Eng. News* **1996**, *74*(25), 43–44; b) Chemical Abstract Service (CAS), Columbus, Ohio, USA, **2007**.

[26] S. Wetzel, W. Wilk, S. Chammaa, B. Sperl, A. G. Roth, A. Yektaoglu, S. Renner, T. Berg, C. Arenz, A. Giannis, T. I. Oprea, D. Rauh, M. Kaiser, H. Waldmann, *Angew. Chem.* **2010**, *122*, 3748–3752; *Angew. Chem. Int. Ed.* **2010**, *49*, 3666–3670.

[27] T. Berg, *ChemBioChem* **2008**, *9*, 2039–2044.

[28] a) K. C. Nicolaou, R. M. Evans, A. J. Roecker, R. Hughes, M. Downes, J. A. Pfefferkorn, *Org. Biomol. Chem.* **2003**, *1*, 908–920; b) K. C. Nicolaou, J. A. Pfefferkorn, H. J. Mitchell, A. J. Roecker, S. Barluenga, G. Q. Cao, R. L. Affleck, J. E. Lillig, *J. Am. Chem. Soc.* **2000**, *122*, 9954–9967.

[29] V. C. Jordan, *Cancer Cell* **2004**, *5*, 207–213.

[30] C. A. de La Lastra, S. Sanchez-Fidalgo, I. Villegas, V. Motilva, *Curr. Pharm. Des.* **2004**, *10*, 3505–3524.

[31] S. Wetzel, PhD Thesis, Univ. Dortmund, **2009**; available at http://eldorado.tu-dortmund.de/handle/03/26470.

[32] M. A. Koch, L. O. Wittenberg, S. Basu, D. A. Jeyaraj, E. Gourzoulidou, K. Reinecke, A. Odermatt, H. Waldmann, *Proc. Natl. Acad. Sci. USA* **2004**, *101*, 16721–16726.

[33] F. J. Dekker, O. Rocks, N. Vartak, S. Menninger, C. Hedberg, R. Balamurugan, S. Wetzel, S. Renner, M. Gerauer, B. Scholermann, M. Rusch, J. W. Kramer, D. Rauh, G. W. Coates, L. Brunsveld, P. I. H. Bastiaens, H. Waldmann, *Nat. Chem. Biol.* **2010**, *6*, 449–456.

[34] a) F. J. Dekker, M. A. Koch, H. Waldmann, *Curr. Opin. Chem. Biol.* **2005**, *9*, 232–239; b) M. A. Koch, H. Waldmann, *Drug Discovery Today* **2005**, *10*, 471–483.

[35] a) H. Waldmann, *Drugs Future* **2009**, *34*, 24–25; b) W. Wilk, T. J. Zimmermann, M. Kaiser, H. Waldmann, *Biol. Chem.* **2010**, *391*, 491–497.

[36] B. M. McArdle, R. J. Quinn, *ChemBioChem* **2007**, *8*, 788–798.

[37] B. M. McArdle, M. R. Campitelli, R. J. Quinn, *J. Nat. Prod.* **2006**, *69*, 14–17.

[38] a) J. P. Nandy, M. Prakesh, S. Khadem, P. T. Reddy, U. Sharma, P. Arya, *Chem. Rev.* **2009**, *109*, 1999–2060; b) T. Lessmann, H. Waldmann, *Chem. Commun.* **2006**, 3380–3389.

[39] a) L. A. Wessjohann, *Curr. Opin. Chem. Biol.* **2000**, *4*, 303–309; b) M. S. Butler, *Nat. Prod. Rep.* **2005**, *22*, 162–195; c) J. W. H. Li, J. C. Vederas, *Science* **2009**, *325*, 161–165; d) J. Y. Ortholand, A. Ganesan, *Curr. Opin. Chem. Biol.* **2004**, *8*, 271–280.

[40] a) D. Brohm, S. Metzger, A. Bhargava, O. Müller, F. Lieb, H. Waldmann, *Angew. Chem.* **2002**, *114*, 319–323; *Angew. Chem. Int. Ed.* **2002**, *41*, 307–311; b) D. Brohm, N. Philippe, S. Metzger, A. Bhargava, O. Müller, F. Lieb, H. Waldmann, *J. Am. Chem. Soc.* **2002**, *124*, 13171–13178.

[41] a) V. Mamane, A. B. Garcia, J. D. Umarye, T. Lessmann, S. Sommer, H. Waldmann, *Tetrahedron* **2007**, *63*, 5754–5767; b) A. B. Garcia, T. Lessmann, J. D. Umarye, V. Mamane, S. Sommer, H. Waldmann, *Chem. Commun.* **2006**, 3868–3870; c) J. D. Umarye, T. Lessmann, A. B. Garcia, V. Mamane, S. Sommer, H. Waldmann, *Chem. Eur. J.* **2007**, *13*, 3305–3319.

[42] F. Perron, K. F. Albizati, *Chem. Rev.* **1989**, *89*, 1617–1661.

[43] D. A. Evans, M. J. Dart, J. L. Duffy, D. L. Rieger, *J. Am. Chem. Soc.* **1995**, *117*, 9073–9074.

[44] a) O. Barun, S. Sommer, H. Waldmann, *Angew. Chem.* **2004**, *116*, 3258–3261; *Angew. Chem. Int. Ed.* **2004**, *43*, 3195–3199; b) O. Barun, K. Kumar, S. Sommer, A. Langerak, T. U. Mayer, O. Müller, H. Waldmann, *Eur. J. Org. Chem.* **2005**, 4773–4788.

[45] I. Paterson, D. Gottschling, D. Menche, *Chem. Commun.* **2005**, 3568–3570.

[46] a) A. Nören-Müller, I. Reis-Corrêa, H. Prinz, C. Rosenbaum, K. Saxena, H. J. Schwalbe, D. Vestweber, G. Cagna, S. Schunk, O. Schwarz, H. Schiewe, H. Waldmann, *Proc. Natl. Acad. Sci. USA* **2006**, *103*, 10606–10611; b) I. R. Corrêa, Jr., A. Noren-Muller, H. D. Ambrosi, S. Jakupovic, K. Saxena, H. Schwalbe, M. Kaiser, H. Waldmann, *Chem. Asian J.* **2007**, *2*, 1109–1126.

[47] A. Nören-Müller, W. Wilk, K. Saxena, H. Schwalbe, M. Kaiser, H. Waldmann, *Angew. Chem.* **2008**, *120*, 6061–6066; *Angew. Chem. Int. Ed.* **2008**, *47*, 5973–5977.

[48] a) A. G. Myers, D. W. Kung, *J. Am. Chem. Soc.* **1999**, *121*, 10828–10829; b) J. F. González, E. de La Cuesta, C. Avendano, *Tetrahedron* **2004**, *60*, 6319–6326; c) C. Chan, R. Heid, S. P. Zheng, J. S. Guo, B. S. Zhou, T. Furuuchi, S. J. Danishefsky, *J. Am. Chem. Soc.* **2005**, *127*, 4596–4598.

[49] S. C. Lee, S. B. Park, *J. Comb. Chem.* **2006**, *8*, 50–57.

[50] D. J. Payne, J. A. Hueso-Rodriguez, H. Boyd, N. O. Concha, C. A. Janson, M. Gilpin, J. H. Bateson, C. Cheever, N. L. Niconovich, S. Pearson, S. Rittenhouse, D. Tew, E. Diez, P. Perez, J. de La Fuente, M. Rees, A. Rivera-Sagredo, *Antimicrob. Agents Chemother.* **2002**, *46*, 1880–1886.

[51] H. Waldmann, V. Khedkar, H. Duckert, M. Schumann, I. M. Oppel, K. Kumar, *Angew. Chem.* **2008**, *120*, 6975–6978; *Angew. Chem. Int. Ed.* **2008**, *47*, 6869–6872.

[52] a) D. G. Rivera, O. E. Vercillo, L. A. Wessjohann, *Org. Biomol. Chem.* **2008**, *6*, 1787–1795; b) L. F. Tietze, M. E. Lieb, *Curr. Opin. Chem. Biol.* **1998**, *2*, 363–371; c) A. Ulaczyk-Lesanko, D. G. Hall, *Curr. Opin. Chem. Biol.* **2005**, *9*, 266–276.

[53] S. Su, J. A. Porco, *J. Am. Chem. Soc.* **2007**, *129*, 7744–7745.

[54] a) C. Bailly, M. Facompre, C. Tardy, C. Mahieu, C. Perez, I. Manzanares, C. Cuevas, *Clin. Cancer Res.* **2003**, *9*, 6112s–6113s; b) E. Marco, W. Laine, C. Tardy, A. Lansiaux, M. Iwao, F. Ishibashi, C. Bailly, F. Gago, *J. Med. Chem.* **2005**, *48*, 3796–3807; c) J. Kluza, M. A. Gallego, A. Loyens, J. C. Beauvillain, J. M. F. Sousa-Faro, C. Cuevas, P. Marchetti, C. Bailly, *Cancer Res.* **2006**, *66*, 3177–3187; d) M. V. R. Reddy, M. R. Rao, D. Rhodes, M. S. T. Hansen, K. Rubins, F. D. Bushman, Y. Venkateswarlu, D. J. Faulkner, *J. Med. Chem.* **1999**, *42*, 1901–1907; e) A. Aubry, X. S. Pan, L. M. Fisher, V. Jarlier, E. Cambau, *Antimicrob. Agents Chemother.* **2004**, *48*, 1281–1288.

[55] H. Waldmann, L. Eberhardt, K. Wittstein, K. Kumar, *Chem. Commun.* **2010**, *46*, 4622–4624.

[56] T. Lessmann, M. G. Leuenberger, S. Menninger, M. Lopez-Canet, O. Müller, S. Hummer, J. Bormann, K. Korn, E. Fava, M.

Zerial, T. U. Mayer, H. Waldmann, *Chem. Biol.* **2007**, *14*, 443–451.

[57] M. A. Sanz, T. Voigt, H. Waldmann, *Adv. Synth. Catal.* **2006**, *348*, 1511–1515.

[58] S. Aoki, K. Higuchi, Y. Ye, R. Satari, M. Kobayashi, *Tetrahedron* **2000**, *56*, 1833–1836.

[59] T. Knoth, K. Warburg, C. Katzka, A. Rai, A. Wolf, A. Brockmeyer, P. Janning, T. F. Reubold, S. Eschenburg, D. J. Manstein, K. Hubel, M. Kaiser, H. Waldmann, *Angew. Chem.* **2009**, *121*, 7376–7381; *Angew. Chem. Int. Ed.* **2009**, *48*, 7240–7245.

[60] O. Rocks, M. Gerauer, N. Vartak, S. Koch, Z. P. Huang, M. Pechlivanis, J. Kuhlmann, L. Brunsveld, A. Chandra, B. Ellinger, H. Waldmann, P. I. H. Bastiaens, *Cell* **2010**, *141*, 458–471.

[61] a) T. Walther, S. Renner, H. Waldmann, H. D. Arndt, *ChemBioChem* **2009**, *10*, 1153–1162; b) T. Walther, H. D. Arndt, H. Waldmann, *Org. Lett.* **2008**, *10*, 3199–3202.

[62] a) H. Waldmann, T. S. Hu, S. Renner, S. Menninger, R. Tannert, T. Oda, H. D. Arndt, *Angew. Chem.* **2008**, *120*, 6573–6577; *Angew. Chem. Int. Ed.* **2008**, *47*, 6473–6477; b) R. Tannert, L. G. Milroy, B. Ellinger, T. S. Hu, H. D. Arndt, H. Waldmann, *J. Am. Chem. Soc.* **2010**, *132*, 3063–3077.

[63] O. Rocks, A. Peyker, M. Kahms, P. J. Verveer, C. Koerner, M. Lumbierres, J. Kuhlmann, H. Waldmann, A. Wittinghofer, P. I. H. Bastiaens, *Science* **2005**, *307*, 1746–1752.

[64] a) L. Bialy, H. Waldmann, *Angew. Chem.* **2005**, *117*, 3880–3906; *Angew. Chem. Int. Ed.* **2005**, *44*, 3814–3839; b) V. V. Vintonyak,

A. P. Antonchick, D. Rauh, H. Waldmann, *Curr. Opin. Chem. Biol.* **2009**, *13*, 272–283.

[65] C. Rosenbaum, P. Baumhof, R. Mazitschek, O. Müller, A. Giannis, H. Waldmann, *Angew. Chem.* **2004**, *116*, 226–230; *Angew. Chem. Int. Ed.* **2004**, *43*, 224–228.

[66] a) M. Wang, *Nutr. Metab.* **2005**, *2*, 3; b) G. Arnaldi, A. Angeli, A. B. Atkinson, X. Bertagna, F. Cavagnini, G. P. Chrousos, G. A. Fava, J. W. Findling, R. C. Gaillard, A. B. Grossman, B. Kola, A. Lacroix, T. Mancini, F. Mantero, J. Newell-Price, L. K. Nieman, N. Sonino, M. L. Vance, A. Giustina, M. Boscaro, *J. Clin. Endocrinol. Metab.* **2003**, *88*, 5593–5602.

[67] a) Y. Kotelevtsev, M. C. Holmes, A. Burchell, P. M. Houston, D. Schmoll, P. Jamieson, R. Best, R. Brown, C. R. W. Edwards, J. R. Seckl, J. J. Mullins, *Proc. Natl. Acad. Sci. USA* **1997**, *94*, 14924–14929; b) N. M. Morton, J. M. Paterson, H. Masuzaki, M. C. Holmes, B. Staels, C. Fievet, B. R. Walker, J. S. Flier, J. J. Mullins, J. R. Seckl, *Diabetes* **2004**, *53*, 931–938.

[68] D. J. Hosfield, Y. Q. Wu, R. J. Skene, M. Hilgers, A. Jennings, G. P. Snell, K. Aertgeerts, *J. Biol. Chem.* **2005**, *280*, 4639–4648.

[69] M. Scheck, M. A. Koch, H. Waldmann, *Tetrahedron* **2008**, *64*, 4792–4802.

[70] H. Gohlke, G. Klebe, *Angew. Chem.* **2002**, *114*, 2764–2798; *Angew. Chem. Int. Ed.* **2002**, *41*, 2644–2676.

[71] a) P. J. Hajduk, *Nat. Chem. Biol.* **2006**, *2*, 658–659; b) P. J. Hajduk, *J. Med. Chem.* **2006**, *49*, 6972–6976.

[72] http://www.ebi.ac.uk/chembldb/index.php.