Algorithm-Aided Enzyme Engineering

Inauguraldissertation

zur

Erlangung des akademischen Grades eines

Doktors der Naturwissenschaften (Dr. rer. nat.)

der

Mathematisch-Naturwissenschaftlichen Fakultät

der

Universität Greifswald

vorgelegt von

David Patsch

Greifswald, Juli 2023

Dekan: Prof. Dr. Gerald Kerth

1. Gutachter: Prof. Dr. Uwe T. Bornscheuer

2. Gutachter: Prof. Dr. Dörte Rother

3. Gutachter: Prof. Dr. Donald Hilvert

Tag der Promotion: 24.10.2023

Table of content

List of abbreviations iii

Scope and outline v

1. Introduction 1
   - 1.1. Current status of protein engineering strategies 1
     - 1.1.1. Directed evolution 1
     - 1.1.2. Rational design 2
     - 1.1.2.1. Rational design strategies 3
     - 1.1.2.1.1. Molecular docking 4
     - 1.1.2.1.2. Consensus approach – evolutionary information 5
     - 1.1.2.1.3. Computational mutation scan 6
     - 1.1.2.1.4. Molecular dynamics for protein design 6
     - 1.1.2.1.5. Stability–function trade-offs 7
     - 1.1.3. Semi-rational design 8
     - 1.1.4. Machine learning-guided directed evolution 11
2. Results 12
   - 2.1. Machine learning-guided directed evolution 12
     - 2.1.1. Overview: Improving enzyme fitness with machine learning (**Article I**) 12
     - 2.1.2. Application: Improving enzyme fitness with machine learning (**Article II**) 13
     - 2.1.3. Reducing enzyme sequence space by predicting and excluding destabilizing mutations (**Article III**) 15
     - 2.1.4. Draft manuscript: Efficient evolution of a Kemp eliminase (**Article IV**) 17
   - 2.3. Asymmetric cation-olefin monocyclization by engineered squalene- hopene cyclases (**Article V**) 19
     - 2.4. Application note: AutoDock Vina plugin for PyMol (**Article VI**) 21
3. Summary 23
4. References 25
5. Author Contributions 34

Articles 36
- Article I 37
- Article II 44
- Article III 123
- Article IV 150
- Article V 167
- Article VI 228

Eigenständigkeitserklärung I
Curriculum Vitae II
List of Publications III
Acknowledgments IV

List of abbreviations

CASP	Critical assessment of protein structure prediction
CAST	Combinatorial active-site saturation test
DFT	Discrete Fourier transform
MD	Molecular dynamics
ML	Machine learning
MSA	Multiple sequence alignment
O_f	Oversampling factor
PCR	Polymerase chain reaction
PROSS	Protein repair one-stop shop
ProSAR	Protein sequence activity relationships
PSSM	Position-specific scoring matrix
REU	Rosetta energy unit
SHC	Squalene-hopene cyclase
SSM	Site saturation mutagenesis

Scope and outline

In their **idealized** forms, enzymes can facilitate complex reactions with extreme specificity and selectivity.[1,2] Additionally, in this imaginative form, they only require mild reaction conditions, resulting in low energy consumption, and they are biodegradable, efficient, reusable, and sustainable.[3] Unfortunately, this idealized form often deviates significantly from reality, where enzymes are more likely to be associated with marginal stability[4] and low reaction rates, leaving them less than desirable for many industrial applications. As such, if we could master the process of engineering the configuration of a protein towards a given task, the implications could be staggering.[2]

This thesis aims to contribute to the process of protein engineering, mainly how computational tools can be used to make the protein engineering process more efficient and accessible.

Article I explores the current state of the art in machine learning-guided directed evolution and serves as a foundation for **Article II**, which is a concrete application of these techniques to an engineering campaign. Despite successfully improving overall activity and selectivity, we also observe limitations and constraints within the methodology. **Article III** then delves into these drawbacks and attempts to lay the foundation for a more generalizable and, more importantly, efficient engineering workflow, balancing the strengths and weaknesses of computational techniques with advances in gene synthesis. We then validated this novel pipeline in **Article IV**, where we show the potential of this methodology. **Article V** describes a more standard protein engineering campaign on squalene-hopene cyclases for potentially interesting products in the flavor and fragrance industry. Lastly, **Article VI** outlines a PyMol plugin for molecular docking.

Article I Improving enzyme fitness with machine learning

D. Patsch, R. Buller, *CHIMIA* **2023**, 77(3), 116. DOI: 10.2533/chimia.2023.116

Machine learning (ML) has permeated most aspects of life, including natural sciences. It holds immense promise for various applications and is expected to do the same for protein engineers. Various attempts have been made to include machine learning techniques in different protein engineering workflows, with varying degrees of success. This article discusses some prominent examples of machine learning-guided directed evolution and how it helped reconfigure various protein properties. Notably, it serves as a foundation for the approach taken in **Article II**. Additionally, the insights gained from previous successful applications of computational methodologies in protein engineering heavily inform the strategy developed in **Article III** and many practical decisions taken in **Article IV**.

Article II **Algorithm-aided engineering of aliphatic halogenase WelO5* for the asymmetric late-stage functionalization of soraphens**

J. Büchler, S. Honda Malca, D. Patsch, M. Voss, N. J. Turner, U. T. Bornscheuer, O. Allemann, C. Le Chapelain, A. Lumbroso, O. Loiseleur, R, Buller, *Nat. Commun.,* **2022,** 13(1), 371. DOI: 10.1038/s41467-022-27999-1

Evaluating mutations at different sites combinatorically quickly leads to libraries of unfathomable dimensions. Even targeting as little as three amino acid positions combinatorically will require screening tens of thousands of variants to cover all possibilities adequately. It is here where machine learning could have a notable impact by learning the underlying sequence-function relationship from a small subset of all possible variants and predicting the properties of unexplored protein sequences. This article studies the halogenase WelO5* and its ability to functionalize the potent anti-fungal agent soraphen A. Based on a partially screened 3-site combinatorial library, we could reliably predict both more active and selective variants for the derivatization of soraphen, reducing the screening burden immensely.

Article III **LibGENiE – A bioinformatic pipeline for the design of information-enriched enzyme libraries**

D. Patsch, M. Eichenberger, M. Voss, U. T. Bornscheuer, R. Buller, submitted to *Comput. Struct. Biotechnol. J.*, **2023**.

It appears that predicting which mutations lead to inactive, misfolded, or insoluble proteins is a much easier task than predicting mutations that improve function. In addition to being a more reliable strategy, this approach is more generalizable, as destabilized enzymes are unlikely to improve any desired trait, such as stability, activity, or selectivity. Combined with recent advances in gene synthesis, this opens a potential path for a robust, generalizable, and efficient protein engineering strategy. **Article III** also includes a web platform, allowing researchers to implement this strategy for their projects quickly.

Article IV **Efficient evolution of a Kemp eliminase**

D. Patsch, M. Voss, T. Schwander, U. T. Bornscheuer, R. Buller, article in preparation, **2023**.

The methodology introduced in **Article III** is applied to evolve the kemp eliminase HG3. This enzyme was previously evolved over 17 rounds of “classical” evolution, which serves as an excellent benchmark for comparison with the pipeline outlined in **Article III**. Notably, HG3.R5, an enzyme with 16 new mutations (compared to the wild-type HG3), emerged after only five rounds, exhibiting a similar activity profile to HG3.17. HG3.R5 shares only one mutation with HG3.17. This article shows that we can navigate the sequence space quickly and efficiently by removing destabilizing mutations and screening the remainder. Interestingly, the evolutionary trajectory between both projects is entirely different, illuminating additional properties of the underlying sequence space.

Article V Asymmetric cation-olefin monocyclization by engineered squalene–hopene cyclases

M. Eichenberger*, S. Hüppi*, D. Patsch*, N. Aeberl, R. Berweger, S. Dossenbach, E. Eichhorn, F. Flachsmann, L. Hortencio, F. Voirol, S. Vollenweider, U. T. Bornscheuer, R. Buller, *Angew. Chem. Int. Ed.*, **2021**, 60(50), 26080–6. DOI: 10.1002/anie.202108037

* equal contribution

Squalene-hopene cyclases (SHCs) are interesting enzymes for industrial cyclic terpene synthesis. **Article V** describes our efforts to gain enantio-complementary access to valuable monocyclic terpenoids with our cooperation partners at Givaudan. Initially, we identified a novel SHC, *Aci*SHC, capable of converting (*E*/*Z*)-geranylacetone to small amounts of monocyclic (*R*)-γ-dihydroionone. We then improved the conversion to this product through process and enzyme optimization to 79 %. The knowledge gained from this reaction allowed us access to the complementary (*S*)-γ-dihydroionone through substrate engineering and further synthetic downstream steps. **Article V** presents an exciting possibility of tuning the absolute configuration of monocyclic products generated enzymatically through substrate engineering.

Article VI AlphaDock – docking plugin

D. Patsch, R. Buller, article in preparation, **2023**.

Molecular docking can help to elucidate and rationalize reaction pathways, protein-ligand interactions, and the impact of amino acid mutations on catalysis. As such, it is a standard and trusted tool in the toolbox of many protein engineers. Recently, AutoDock Vina, one of the most popular and widely used molecular docking programs, was updated and introduced various new features. **Article IV** describes AlphaDock, a PyMol plugin that aims to give researchers access to these new features and help with reproducibility and installation.

1. Introduction

1.1. Current status of protein engineering strategies

1.1.1. Directed evolution

Directed evolution has become a powerful technique for protein engineering over the years[5] and is commonly used to configure critical protein properties, such as activity, thermo-/solvent stability, selectivity, and specificity.[6,7] It mimics the natural selection process by subjecting a protein to cycles of generating genetic diversity, followed by selection for improved traits.[5] However, unlike nature which selects for survival or reproduction, directed evolution creates diversity much more aggressively and only carries those variants with improved biological function (fitness) to the next generation.[8] This approach was incredibly influential in protein engineering and was awarded the 2018 Nobel Prize for chemistry.[9] Genetic diversification is most commonly achieved by one of two approaches: random mutagenesis (error-prone PCR) and recombination of related sequences (gene shuffling)[8], both of which are approachable to most researchers. The random nature of this diversification leads to variations that are distant from the active site (which, conversely, usually is the focus of most targeted approaches because it accommodates the most beneficial mutations)[10,11] to be discovered.[12] Additionally, no structural information about the protein is required. The lack of required information about the system makes directed evolution an attractive starting point for many engineering campaigns.

Nevertheless, directed evolution can only barely scratch the surface of all possible enzyme variations. There are thousands of ways to alter a protein sequence by one change, millions to alter it by two, and so on,[13] making it impossible to search this sequence space exhaustively. Notably, the random nature of traditional directed evolution, in which this space is explored, can also be seen as one of its drawbacks. Most changes introduced into a protein are either neutral or unfavorable,[14] leading to an inefficient sampling of the sequence space. To make matters worse, the impact of individual mutations is often tiny, facilitating the need for multiple rounds of evolution to reach acceptable fitness levels.[12,15] This sampling (or screening) usually represents the main bottleneck in protein engineering campaigns, as only a limited number of variants can be evaluated feasibly (typically in the range of 10^3-10^4).

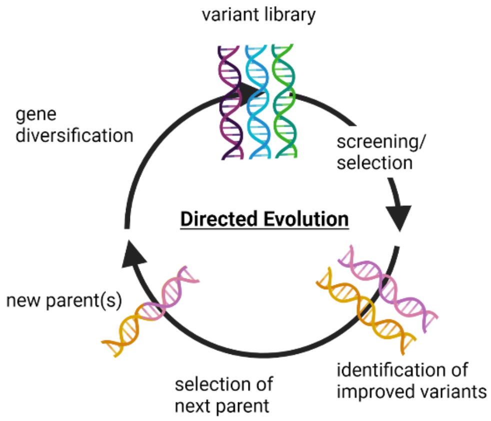


Figure 1: Principle of directed evolution. Directed evolution mimics the natural selection process by subjecting a protein to cycles of generating genetic diversity, followed by screening for improved traits. Gene diversification is most often achieved through error-prone PCR or gene shuffling. The resulting variant library is then screened. This screening represents the main bottleneck in directed evolution, as the theoretical size of the generated variant library is astronomical. Improved variants are then characterized to select a parent for the next round. This iterative process runs until no further improvements are achieved or the project goals are met. Image from Patsch & Buller, 2023.[16]

1.1.2. Rational design

Rational design emerged as an alternative to directed evolution.[17] It aims to limit the screening effort to only a few amino acid substitutions based on an intimate knowledge of the function or structure of a protein.[18] This can significantly reduce library size and, by extension, screening efforts. Rational design is particularly appealing when no high throughput assay is available.[12]

For example, information from multiple sequence alignments (MSA)[19,20] or structure-based computational design,[21,22] and molecular dynamics (MD)[23] simulations may be employed to increase enzyme stability and functional expression. These techniques are a cornerstone of many protein engineering strategies. For this reason, they will be discussed in more detail in separate sections below.

However, it must be emphasized that these computational techniques are not a silver bullet. A high prediction accuracy is required since even a single deleterious mutation can invalidate the engineering effort.[24] Achieving such an accuracy remains a significant challenge for specific tasks such as altering or inverting enantioselectivity.[17]

Additionally, researchers may find rational design less approachable and generalizable than directed evolution. A thorough computational background is required on top of the mandatory expertise in the biological/chemical aspects of protein engineering. It can be challenging to rationalize which sites, specific residues, or combinations of mutations should be evaluated.

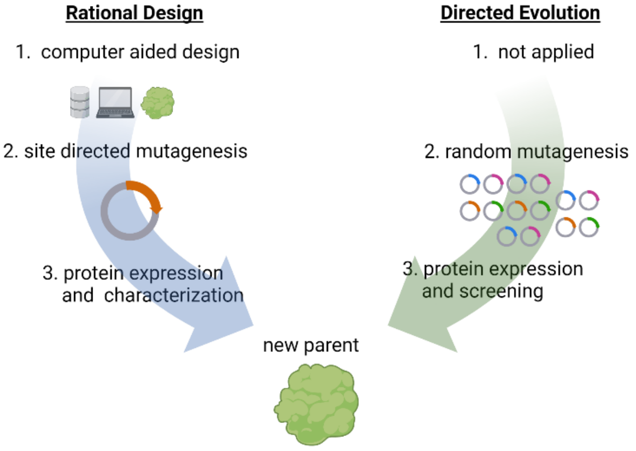

Rational Design
1. computer aided design
2. site directed mutagenesis
3. protein expression and characterization
Directed Evolution
1. not applied
2. random mutagenesis
3. protein expression and screening
new parent

Figure 2: Rational design and directed evolution. In rational design, information from various sources is used to design specific variants. Diversity is created significantly more randomly in directed evolution, resulting in much larger libraries. Rational design can be much more appealing when no high-throughput assay is available.[25] *However, depending on the task, individual mutations might not contribute significantly to fitness. This requires the introduction of multiple mutations, which requires exceptional prediction accuracy. Image adapted from Dvorak et al., 2007.*[26]

1.1.2.1. Rational design strategies

The optimal procedure for selecting potentially beneficial mutations or variants depends on the characteristic to be configured. Targets related to catalysis, such as selectivity, specificity, or activity, require different approaches than targets related to expression or stability. The former is often related to residues that influence substrate binding, stabilize transition states, or facilitate product release.[27,28] These residues can be identified, for example, with tools to predict/analyze molecular docking,[29] study receptor-ligand interaction networks,[30] or tools to study the binding pocket and its access tunnels.[31–34] Techniques for improving stability or expression (which can sometimes be hard to separate) include consensus design, co-evolution, rigidifying flexible protein areas, and redesigning surface charges.[35–39]

These strategies vary in complexity and difficulty, but they all aim to identify specific residues for experimental evaluation. However, there are no definitive guidelines for their applicability, often leading researchers to rely on heuristics when balancing the diverse goals of an engineering campaign. Streamlining the site selection process could significantly enhance the accessibility and efficiency of protein engineering. The upcoming sections will concisely introduce commonly used rational design tools, focusing on those relevant to this thesis and their potential applications in different scenarios. **Article III** and **Article IV** will then focus on how this array of tools and options could be more consistently integrated into protein engineering workflows with differing objectives.

1.1.2.1.1. Molecular docking

Molecular docking tools are routinely used to predict the orientation of a ligand when bound to an enzyme. Such predictions can be helpful in various applications; For example, millions of compounds can be docked against a receptor to select potential lead candidates for treatments from enormous chemical libraries. This has led to the identification of STX-0119 for lymphoma,[40] novel human histamine H4 receptor ligands,[41] and Pim-1 kinase inhibitors,[42] as well as potential springboards for therapeutic designs toward SARS-CoV-2,[43,44] among many others.

In rational design, docking can give insights on residues that might interfere with a ligand of interest or hint at which ones should be replaced to improve binding or better stabilize transition states (in combination with interaction network analysis). Understanding which residues are close to the bound ligand can be valuable by itself. Those positions are the target of many rational/semi-rational protein engineering strategies considering the active site often harbors most of the beneficial mutations.[45] Molecular docking is also very commonly used at the end of protein engineering campaigns to rationalize differences in catalytic properties (for example, why the improved variant is more selective than the wild-type).

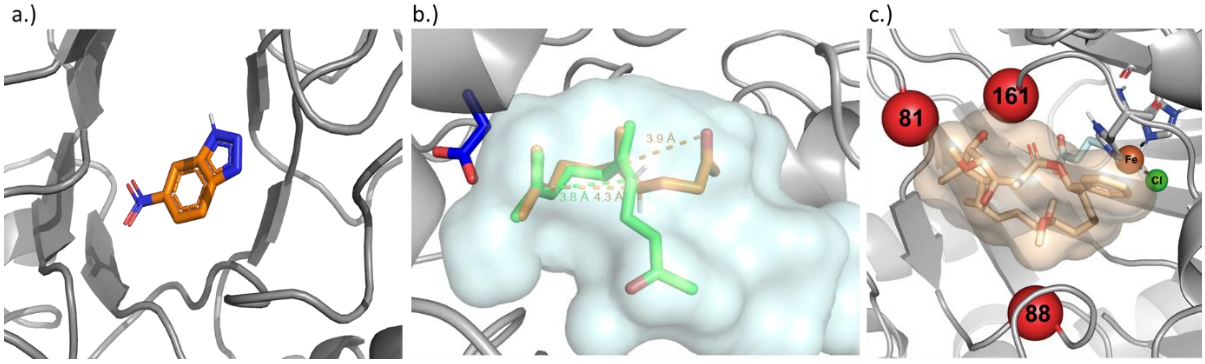


Figure 3: Examples of applications of molecular docking in this thesis. a.) Kemp Eliminase HG3.17 docked with the transition state analogue 6-nitrobenzotriazole.[15] *b.) Docking of the substrate isomers (E/Z)-geranylacetone in the homology model of AciSHC.*[46] *c.) Docking of soraphen A into a model of variant WelO5*.*[47]

As docking is such a vital aspect of drug discovery and protein engineering, over the years, more than 60 tools have been developed for this purpose.[48] Docking software generally uses a scoring function to quantify the protein-ligand complex's chemical potentials, electrostatic potentials, and shape.[49] In most cases, this consists of sampling different conformations of a flexible receptor in a rigid receptor, then optimizing this conformation based on the tool-specific scoring function.[48]

One of the fastest and most widely used tools for this task is AutoDock-Vina.[50] It was initially developed and published by Dr. Oleg Trott in 2010 but was recently improved and expanded by the Forli lab at The Scripps Research Institute.[29] The researchers added various new features that make AutoDock-Vina more attractive.

These features include:

- Multiple ligand docking: for example, ligand and cofactor.
- Flexible residue docking: rather than treating the receptor as a completely rigid object, this allows for selected residues to be moved. To be more specific, it allows for different rotamers of selected residues. The backbone (C-alpha atoms) stays in place.
- Hydrated docking: to model water involved in protein-ligand interactions, which is often neglected.
- New scoring functions: allow the exploration of the energy landscape with different search algorithms.

Notably, the new AutoDock-Vina also exposes many new options and configurations to tailor the docking process to the users' needs.

Article VI will introduce AlphaDock, a PyMol plugin developed within the scope of this thesis that aims to make the new AutoDock-Vina features more accessible to researchers.

1.1.2.1.2. Consensus approach – evolutionary information

Consensus-based protein engineering assumes that, on average, the consensus amino acid (the most conserved residue in a multiple sequence alignment) contributes more to protein fitness than the other possible amino acid substitutions at this site.[51] The intuition behind this assumption is that deleterious or disadvantageous mutations tend to be purged by evolution.[24] Interestingly, these changes often positively affect specific protein properties, particularly stability. As with other protein engineering techniques, there is no clearly defined way to apply this approach. The basic principle that the consensus residue (the most common residue in the evolutionary context) replaces the wild-type residue stays the same. However, the thresholds that define "consensus" do not. A popular web server for the identification of hot spots, aptly titled HotSpot Wizard,[52,53] suggests the following: Consensus design can be implemented as either frequency or majority based. In frequency-based design, the wild-type residue is replaced by the consensus if the consensus residue is present in at least 50 % of all analyzed sequences. The criterion for majority based is different. Here, the wild-type is already replaced if the consensus value is 40 %; however, the consensus residue must also be more than five times as frequent as the wild-type. The Protein Repair One-Stop Shop (PROSS) web server[24] applies a different heuristic. Here, a position-specific substitution matrix (PSSM) is computed,[54] and only residues with a favorable PSSM score (>0) are considered. The PSSM represents the log probability of observing any given amino acid at any position in the protein.[24]

The source of the MSAs that serve as the foundation for consensus design also varies. A good starting point might be sequence databases such as Pfam[55] and Superfamily[56] or querying more extensive databases such as UniProtKB/Swiss-Prot,[57] the Protein Data Bank (PDB), and NCBI.[58]

The variation of ways to source evolutionary information and the differences in evaluating it will inevitably lead to different results. However, as large-scale validation datasets do not exist, proving the optimal strategy is impossible. Nevertheless, a strategy derived from previous experiences is often enough to provide satisfactory results and improve a desired trait.[24,52,53]

1.1.2.1.3. Computational mutation scan

Another common strategy in protein engineering is evaluating the effect of mutations computationally. Force field calculations can indicate whether any given mutation positively or negatively influences a protein characteristic.[53] Similarly to consensus-based design, this is mainly used to target stability/expression. The intuition here is that flexible residues have few interactions with their neighbors, so replacing them with different residues that form contacts can lead to more stable proteins.[59–61]

The impact of amino acid substitutions, single point or multiple point mutants, are quantified by evaluating the difference in predicted free energy between the wild-type and mutant ($\Delta\Delta G$).[62] This requires efficient conformational sampling methods to create the mutant and an accurate free energy function to score it.[63] Some of the most popular modeling tools for these tasks in the academic community include FoldX[64] and Rosetta.[64] However, different tools might be required for different applications. Depending on computational resources and available licenses, the researcher/practitioner might reach for alternatives, of which there are plenty.

1.1.2.1.4. Molecular dynamics for protein design

Molecular dynamics (MD) is a critical tool in protein engineering. It is a way to simulate how atoms and molecules move and behave over time. This is achieved by solving Newton's equations of motion in a system of moving and interplaying particles. A physics-based potential energy function or force field calculates their properties and forces.[23,65] These force fields are either generated empirically (from experiments) or calculated from quantum mechanics or a combination.[66]

Researchers routinely rely on MD to craft more efficient biocatalysts and therapeutics. Studying the behavior of a protein over time develops the understanding of protein chemistry and folding,[23] which in turn builds a stronger intuition of how a protein might function.

MD can easily be combined with other tools as well. For example, most docking tools treat the receptor as a rigid object - a simplification - which does not reflect reality. The structural flexibility of receptors should ideally be considered during modeling,[67] as it plays a vital role in protein-ligand complex formation.

This is similar to moving from the key-lock representation to the induced-fit model.[68] Conceptually this can be achieved by an iterative process of docking a flexible ligand into a rigid receptor, then performing a short MD simulation on the resulting complex. This additional malleability allows the receptor to "fit" to the introduced ligand. Then the ligand is docked again into the adjusted protein.[69] Notably, improved accuracy comes at the cost of the docking procedure being much more computationally expensive. Additionally, the process does not necessarily have to be performed iteratively. Docking results can already be improved significantly by running short MD simulations on the output from AutoDock Vina.[70]

Molecular dynamics can also be of great use when designing proteins for higher stability, expression, or resistance to harsh environments. The strategy is called "rigidification of flexible sites",[37] which can be identified through MD simulations. In combination with ΔΔG calculations and evaluating the evolutionary context, the researcher can identify sites of interest and which substitutions at these sites they should consider.

The most significant limitation of MD is its computational cost. To employ MD for studying protein folding and flexibility, or even binding/unbinding events, the simulation has to be run for periods upward of 20-500 ns.[71] Such timeframes can be prohibitively expensive for most researchers. On top of the computational cost, the computational complexity of setting up the system, with appropriate forcefields and parameters, must not be underestimated. For this reason, researchers often turn to B-factors as an alternative metric for protein flexibility. B-factors are an experimental metric for protein flexibility and occur during crystallization due to X-ray scattering from thermal motion.[72] If ions and small molecules related to the protein were co-crystallized, B-factors give a reliable and easy estimate of flexible regions within the macromolecule.

1.1.2.1.5. Stability–function trade-offs

In 2014, Hyun June Park and colleagues published a study on a computational design strategy for *Candida antarctica* lipase B (CalB).[73] The researchers took a very rational approach toward protein design. Initially, they performed MD simulations at increasing temperatures (300 K, 330 K, 360 K, and 400 K). Doing so led to identifying seven residues that fluctuated the most as the simulation temperature increased. These seven sites were then further investigated with the Rosetta tool to predict specific residues that might improve stability over the wild-type (see chapter 1.1.2.1.3.), resulting in three specific mutants: A146D, T158S, and A251E. Then, they performed MD simulations and physical stability experiments on these newly created variants. Regarding their simulation and stability profile, A146D and T158S behave similarly/worse to the wild-type. However, variant A251E does not. It shows less flexibility during the MD run and improved stability characteristics compared to the wild-type. Their experimental setup measured residual activity after incubation at 50 °C for four hours.

Interestingly, when they measured specific activity, the more stable variant A251E only exhibited 50 % of the performance of the wild-type. In contrast, the slightly less stable (compared to the wild-type) variant T158S exhibits significantly higher specific activity than the wild-type under native conditions.

This study highlights an important trade-off: as the thermodynamic stability increased, the relative activity decreased, and vice versa. However, this is not always the case. There have been reports of negative[74–77] and positive correlations[78–80] between rigidity (thermodynamic stability) and function. The critical consequence of these conflicting results is that the relationship between distinct protein traits might be challenging to generalize, as different proteins behave differently in different situations.[23]

Multiple goals must be balanced during a protein engineering campaign, and different techniques might be required depending on the goal. In addition, the application and combination of these techniques are often based on heuristics, and no clear guidelines exist. This can result in unnecessarily complex and inefficient processes. **Articles III and IV** focus on how computational tools can be more consistently integrated into protein engineering workflows with the aim of reducing cost and complexity.

1.1.3. Semi-rational design

The design tools for directed evolution have become increasingly diverse and range from purely random[81] to highly rational.[82] While we separate the techniques in name, they often overlap, and a clear distinction becomes challenging to draw in practice.[25] The vastness of the combinatorial sequence space requires some rational input, and the limitations and inaccuracies of computational techniques often require additional variants to be screened. This middle ground, called semi-rational design, combines elements of rational design and directed evolution to create smaller, more focused libraries of higher quality.[83,84] Ideally, such a combination compensates for individual techniques' shortcomings and bolsters their strengths. Researchers narrow the sequence space to potential hotspots based on information from sources such as docking, machine learning, phylogeny, the 3D structure, function, or previous knowledge (such as mutational data).[17,83] Rather than only a handful of variants, researchers use this information to construct focused libraries (also called smart libraries) ranging in size from ~200-2000 samples. The exact number is chosen based on practical considerations, such as the capabilities of the required analytical systems and screening assays.[85]

Combining computational and random techniques in this way addresses some significant drawbacks of the individual strategies. On the one hand, focused libraries lead to a more efficient sampling of the sequence space, resulting in a lower screening burden than traditional directed evolution.[86,87] On the other hand, evaluating more samples than in rational design allows for more leniency with respect to computational limitations and inaccuracies.

Although semi-rational design has some attractive advantages, researchers may find it even less approachable than rational design. The desire to screen more than a handful of variants also raises practical aspects to consider, such as, for example, the physical construction of the designed smart libraries.

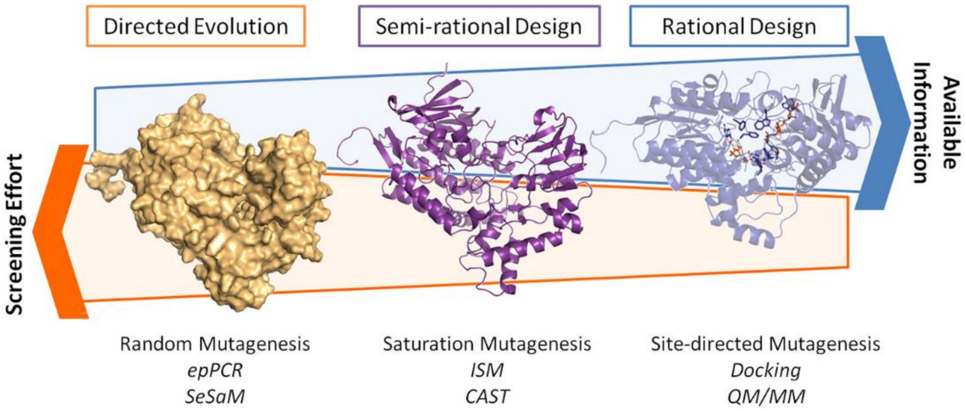


Figure 4: Protein engineering strategies, sorted by screening effort. In traditional evolution, the sequence space is explored randomly. In rational design, on the other hand, specific variants for evaluation are carefully planned and designed. The factors that define which strategy to use include, among others, the available information and screening capabilities. Image from Balke et al., 2017.[88]

The most prominent example of semi-rational design is the combinatorial active site saturation test (CAST). Here, the protein crystal structure or a protein homology model is used to identify residues in the binding pocket (in the simplest case – the technique has been extensively adapted to various problems).[61] A few selected sites are then randomized, individually or in combination, relying, for example, on the popular QuikChange protocol, resulting in economic libraries of high quality.[89] Evaluating combinations combinatorically might reveal synergistic effects that would have been missed otherwise. However, such an approach introduces a whole new set of challenges; For one, the library size increases exponentially with the number of sites chosen (three sites: 8 000 possible combinations, four sites: 160 000 combinations); for another, the issue of oversampling cannot be neglected. As the number of mutants increases, so does the probability of generating and screening a duplicate variant. As a result, an increased number of variants need to be screened to obtain a particular library coverage. Assuming each sequence occurs with equal probability, the required "oversampling factor" (O_f) can be calculated as described in Equation (1).[90]

$$O = -\ln(1-P_i) \tag{1}$$

Where P_i describes the probability of a particular sequence occurring in the library.

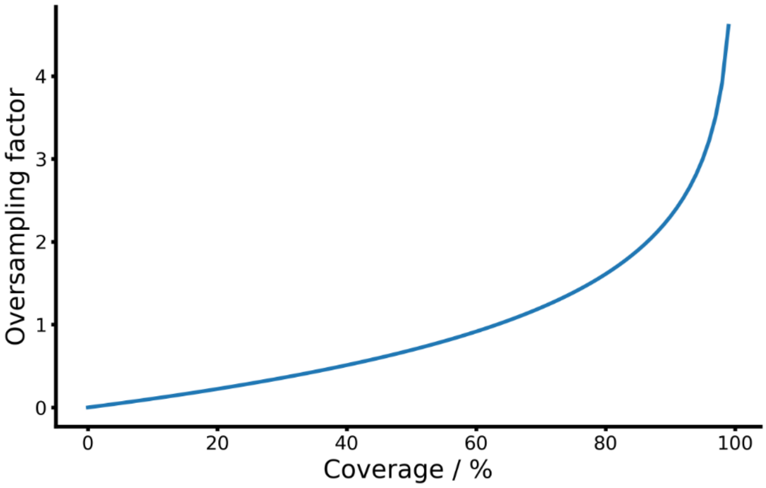


Figure 5: Calculating the oversampling factor as a function of percent coverage as described in Eq. (1).

Consequently, to achieve 95 % coverage, approximately three times the number of variants must be screened (Figure 5). This oversampling represents a major bottleneck in protein engineering and multiplies the already significant screening effort.

Machine learning could provide a way to bridge this gap by learning sequence-function relationships on a smaller subset of the entire library and making predictions on the remainder, sidestepping the issue of oversampling and simultaneously reducing the screening burden. In contrast to traditional directed evolution, which discards information about everything but the most beneficial mutations, machine learning techniques might be able to use this data to speed up the evolution process by learning a function representing the underlying protein landscape from a set of sequence-fitness pairs. With this function, additional variants can be "screened" *in-silico*, allowing variants to be evaluated at a scale/pace that cannot be accomplished with wet-lab experiments alone.[91] The potential benefits of ML make it an attractive research objective, and multiple attempts to apply it to protein engineering have been made, including efforts to increase enzyme activity[92,93] and stability[94,95].

1.1.4. Machine learning-guided directed evolution

Machine learning already affects many areas of our daily lives, from translating languages[96] to suggesting movies we might like,[97] and is rapidly making its way into traditional life sciences as well. The most notable example is AlphaFold 2, which swept away its competition in the 14th Critical Assessment of Protein Structure Prediction Challenge (CASP14).[98] For the first time, the winning solution demonstrates competitive accuracy with experimentally determined structures, even when no similar structures are known.[99] This is undoubtedly an astounding achievement; some even consider it a solution to protein-folding, a decades-old research problem in biology.[100] AlphaFold 2 has given rise to other algorithms that reduce the time required to predict a 3D structure with atomic accuracy from a protein sequence even further, down to only a few seconds.[101] Given what has happened in the past few years, it is reasonable to assume that other areas of protein engineering also stand to gain significantly from machine learning.

High-accuracy and easily accessible protein structures from models such as AlphaFold 2 are already helpful to protein engineers. However, machine learning could have the most beneficial impact by reducing the screening burden. Ideally, the sequence-function data of variants screened in a directed evolution campaign could be used to predict which variants to evaluate next. The only additional costs this strategy would incur are sequencing and the required computing power, which are constantly getting more affordable.

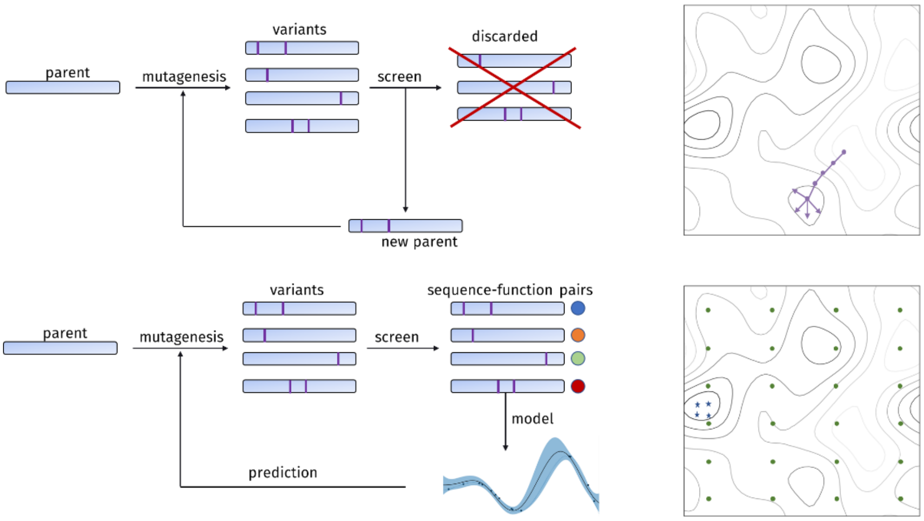


Figure 6: Directed evolution with and without machine learning. Image adapted from Yang et al., 2019.[102]

Machine learning has already found application in protein engineering for various problems, from improving enantioselectivity[17,103] to enzyme activity[47] and stability.[104] Although the targets may vary, the process is usually similar. Initially, a part of the sequence space is screened to create a dataset of annotated function pairs. Then, these pairs are first represented as a vector before training a model and relying on that to guide further exploration. **Article I** reviews previous applications of machine learning in protein engineering, and **Article II** applies these techniques to tailor a halogenase for macrolide derivatization.

2. Results

2.1. Machine learning-guided directed evolution

2.1.1. Overview: Improving enzyme fitness with machine learning (**Article I**)

Using machine learning techniques to reduce the astronomically large sequence space constituted by all possible amino acid combinations is an attractive proposition. As such, multiple attempts in this regard have been made. While these attempts vary wildly in their application and often rely on different techniques, models and encoding strategies, they also highlight some important commonalities.

First, it is essential to realize that not all facets of protein engineering can benefit equally from machine learning. Most importantly, machine learning algorithms work best with clean and reproducible data (garbage in, garbage out),[105] which cannot always be asserted in biological systems. Additionally, the costs and especially time requirements of sequencing all variants can be crucial in opting for different optimization strategies. Machine learning might also not be advisable in an ultra-high-throughput system, where the overhead of sequencing and predicting far outweighs the cost of screening. The same could be said about the opposite, very low throughput approaches, as only a handful of samples will not produce reliable predictions. However, from experience, most engineering campaigns do not operate in these extremes and could accommodate ML into their workflows.

Once the decision to rely on machine learning to guide directed evolution has been made, one has to decide how to **i**) represent sequences and train the sequence-function model and **ii**) which predicted sequences should be evaluated in the laboratory. There exists a significant amount of diversity within each of these points. For example, a protein can be encoded numerically in various ways. Feng et al.[103] represent each amino acid by one-hot encoding (each site consists of a 20-dimensional vector of zeros, except for the position of the specific residue at that position, which is represented as a one). Rather than a simple vector of ones and zeros, proteins can also be encoded by their physicochemical and biochemical properties.[106] Ideally, introducing such information can help the models make better decisions with fewer training data points. Cadet et al.[93] even went a step further and processed the obtained numerical sequences by means of Discrete Fourier Transform (DFT). The researchers intend to process the protein signals to reveal additional information embedded in them.[93,107] Recently, advances in unsupervised learning have made it possible to represent the protein as a whole, not just by its individual amino acids. These representations contain information from physicochemical properties to remote homology and structural components, allowing for improved predictions across various tasks.[108]

Like the diverse ways a protein can be encoded, multiple techniques have been applied to learn sequence-fitness relationships with varying complexity. Interestingly, some of the most prominent examples of protein engineering rely on the analysis of protein sequence activity relationships (ProSAR) developed by the US-based company Codexis.[104,109] This technique is based on simple one-hot encoding and statistical analysis through linear regression. ProSAR's objective is not always to select the best variant in each round but to rapidly identify favorable recombination mutations to attain fitness targets. The researchers at Codexis attribute the decision to move forward with a decent variant, instead of being tied down with an exhaustive search for the optimal enzyme, as the key to their successful evolution campaigns.[104]

The Codexis examples highlight the critical practical circumstances behind algorithm-aided directed evolution. Considering the entire process is crucial when designing novel algorithms and strategies. What is the optimal way to introduce machine learning into protein engineering workflows? What areas can benefit the most? What model or method is the best? How many predictions should be evaluated?

At the time of writing, these questions are hard to answer generally. No clearly defined benchmarks exist to compare different machine learning-guided directed evolution techniques. Different protein encodings, machine learning models, hyperparameters, etc., will have to be tested and validated on each task to maximize predictive accuracy. More protein engineering examples that rely on machine learning are required to validate algorithms on more than a handful of datasets.[47,110]

2.1.2. Application: Improving enzyme fitness with machine learning (**Article II**)

We sought to explore the feasibility of machine learning to configure enzymatic activity and regioselectivity. Towards this end, in **Article II**, we first studied the halogenase WelO5*[111] and its ability to halogenate soraphen A selectively.[47] While the wild-type enzyme is incapable of doing so, more promiscuous WelO5* variants were constructed in previous studies, and an engineered halogenase showing activity toward soraphen A, producing two products **1a** and

1b (Figure 7a), was identified. We selected three sites (V81/A88/I161) for complete combinatorial randomization based on these previous experiments and additional docking studies. These three sites already lead to a library size of 8.000 (20^3); however, considering the required oversampling factor of three,[90] roughly ~27.000 variants (assuming only 20 codons at each site) need to be screened to achieve 95 % library coverage. As such, we determined this to be an ideal application for machine learning-guided directed evolution.

A total of 504 unique variants were confirmed experimentally, and each sequence was associated with activity data. This corresponds to a library coverage of 6.3 %. Different encoding strategies and machine learning techniques were explored and validated through a 10-fold cross-validation scheme to assert generalizability and maximize predictive accuracy. In the final process, each WelO5* mutant in the library was represented as a vector, created by concatenating each residue's physicochemical and biochemical properties at each of the three sites. Predictions towards activity were made with Gaussian processes. The encoding and model selection scheme was similar to previously reported approaches.[112] The best predicted variants towards activity and regioselectivity were then experimentally validated. We were pleasantly surprised to discover that all seven variants anticipated to have increased activity functioned well, with four halogenases even beating the previous best variant (as shown in Figure 7). Furthermore, the variants predicted to have higher selectivity displayed the desired enzyme characteristic. Specifically, seven of eight produced halogenases showed high selectivity toward the chlorinated soraphen regioisomer **1b**, while the variant "AHG" showed absolute regio-selectivity and doubled activity compared to the previous best **1b-**producing variant.

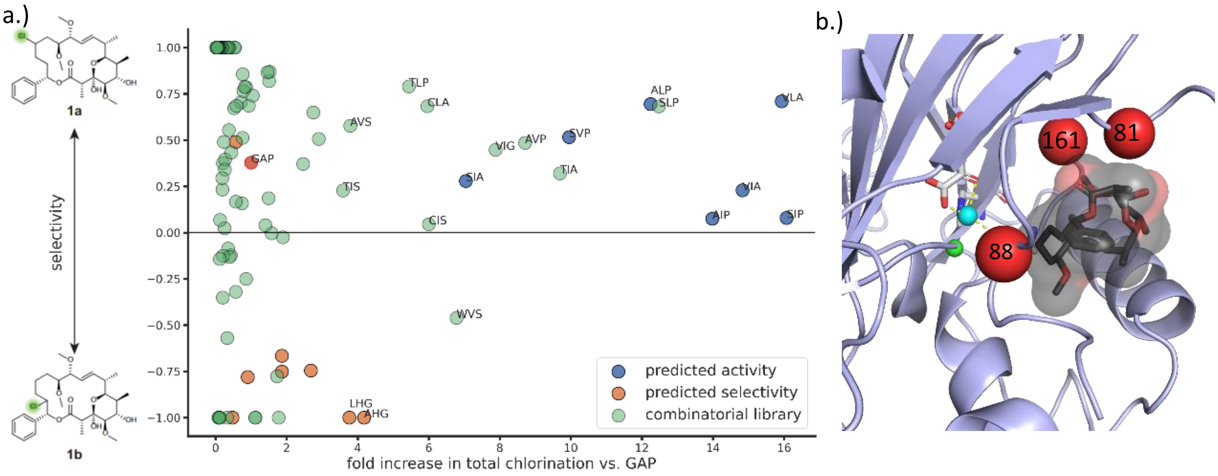


Figure 7: a.) The experimentally measured activity and regioselectivity data from the three-site WelO5 combinatorial library (green). The predicted variants towards activity are highlighted in blue and mutants predicted towards selectivity in orange. The y-axis represents chlorination regioselectivity. b.) Soraphen A, docked in a model variant of WelO5* V81G/I161P. Image from Patsch & Buller, 2023.*[16]

2.1.3. Reducing enzyme sequence space by predicting and excluding destabilizing mutations (**Article III**)

Machine learning models can predict patterns and guess possible outcomes based on the training data. However, the fitness landscape is not always rugged, and mutations can behave in foreseeable ways – by combining additively to produce more active variants.[113] An essential aspect of a machine learning-guided approach would be the ability to pick up on epistatic effects that might not be present in the training data. Specifically, this refers to epistatic effects that improve fitness and could not have been identified typically, particularly cases such as combining two mutations that individually perform poorly to produce a variant with notably better fitness. Doing so is far from trivial. It will require researchers to show that it is possible to predict epistasis in the first place and that this approach can be generalized to other systems. In our research, we only observed such an epistatic interaction once out of thousands of data points, and interactions presented in the literature can often also be found through simple addition.

It is essential to ask what we can predict, how these predictions can be validated, how our strategies can generalize, and how we can justify the added cost of computational techniques. This reasoning led to **Article III**. Various computational studies revealed that predicting which mutations are destabilizing (leading to a decrease in function) is much easier and more reliable than predicting those mutations that improve function.[24] This is an important distinction and allows for different techniques to be used in ways that might not be immediately obvious, such as enabling the use of stability predictions for activity optimization.

For instance, Codexis evolved a carbonic anhydrase towards improved activity at higher temperatures by saturating all non-catalytic residues in the first round.[104] Through this initial screening, they identified 84 unique mutations that performed better than the wild-type under their screening conditions. We calculated the predicted ΔΔG values for all single-point mutants and noticed that most of these improved variants were within the top 60 % of predicted ΔΔG values (cartesian ΔΔG protocol),[64] implying that a significant portion of the screening space could have been excluded computationally (Figure 8b).

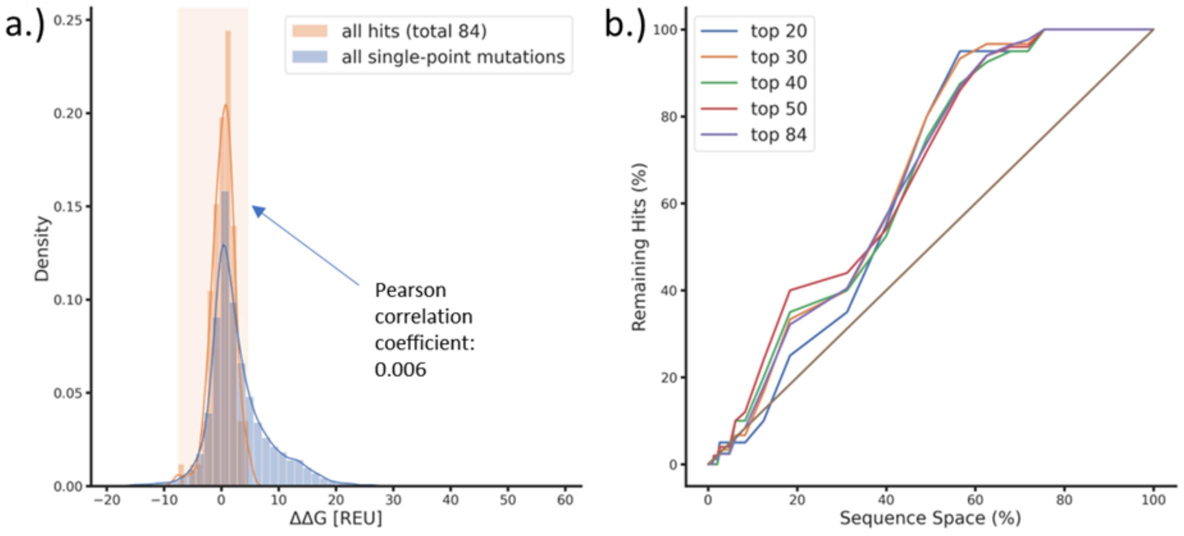


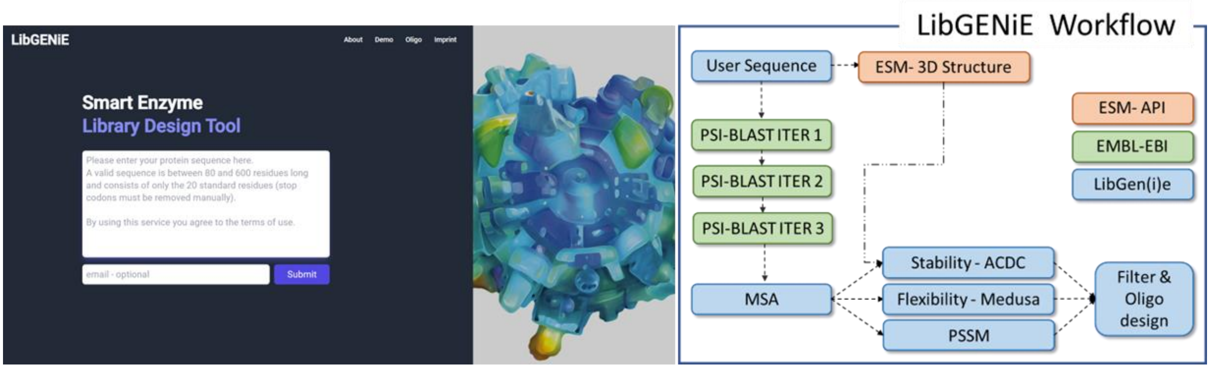
Figure 8: a.) Predicted ΔΔG values of single-point mutants of a carbonic anhydrase.[104] *Lower values indicate higher predicted stability. The blue density plot corresponds to all possible single-point mutants, whereas the orange density plot/area shows the distribution of all 84 identified beneficial mutations. b.) The same data, but visualized as a line chart. Subplot* ***a.)*** *indicates that a part of the sequence space could have been excluded by predicting variants that destabilize the protein. Subplot* ***b.)*** *depicts how reducing the sequence space (x-axis) by removing destabilizing mutations affects the remaining hits. For example, if the most stabilizing 40 % of sequences are removed, most (> 90 %) of the initial 84 mutations remain. Image from draft* ***Article III****.*

Notably, predicted ΔΔG values become much less informative below a certain exclusion threshold. Fold improvement over wild-type is not correlated to predicted ΔΔG values in the ΔΔG range where hits were discovered (-7.5 to 4.7 rosetta energy units - REU). We also observed this pattern, that improved variants are not predicted to be strongly stabilizing in different systems with different objectives.

This finding by itself is not necessarily useful yet. Filtering the sequence space by removing destabilizing mutations results in libraries that are way too diversified to be economically covered through traditional degenerated primers. One possible solution to the problem of constructing complex custom libraries could be micro-array-synthesized oligonucleotides, also knowns "oligo-pools".[114] The potential upsides and applications of oligo-pools make them incredibly interesting for protein engineers in other aspects besides constructing pre-filtered libraries. For example, oligo-pools could significantly impact the application of common strategies such as alanine scans,[115,116] or large-scale saturation projects (SSM on every other site, n-th sphere,[117] full non-catalytic).[104]

However, we noticed relatively high error rates in preliminary experiments with these oligo-pools. Only roughly 52 % of all sequences exhibit desired mutations, with the remaining 48 % being split between wild-type and multiple-point mutants. Such rates are also within what is expected from the literature.[118–120] We slightly improved this in **Article III** by optimizing the PCR protocol, resulting in roughly 60 % of all sequences displaying the correct mutations.

Removing destabilizing mutations from enzyme libraries and screening the rest has some interesting properties; however, computational resources and complexities might limit its accessibility. As such, **Article III** also introduces LibGENiE, a web platform to create smart libraries and design oligo pools. By providing common protein properties, such as stability, evolutionary context, and flexibility, users can reduce their sequence space. Additionally, the website makes it easy for users to design complex oligo libraries to create the filtered library.



Please enter your protein sequence here.
A valid sequence is between 80 and 600 residues long and consists of only the 20 standard residues (stop codons must be removed manually).

By using this service you agree to the terms of use.

*Figure 9: a.) Schematic overview of the LibGENiE workflow. An MSA and a 3D structure are created from the user input sequence through publicly accessible APIs. Then, different tools are used to predict various protein properties. b.) LibGENiE will automatically split the gene into fragments of a desired length and design all possible single-point mutations as well as pool amplification primers. Image from draft **Article III.***

2.1.4. Draft manuscript: Efficient evolution of a Kemp eliminase (**Article IV**)

Article IV focuses on the concrete application of the methodology outlined in **Article III**. We observed that various tools and measures are required in a standard protein engineering campaign, depending on the desired protein characteristic to evolve. Even with a thorough computational background (just one facet of the entire process) and understanding of the system, defining a strategy and which sites or residues to reconfigure can be challenging.

However, we observe that a fitness decrease also accompanies mutations destabilizing the protein. This way, the same strategy – removing undesired variants – can be applied to multiple tasks. By relying on external tools, such as the website introduced in **Article III**, it becomes possible to design libraries and the required oligo sequences to construct them in a few hours.

To better understand the benefits and potential downfalls of the filter/oligo-based strategy, we sought to re-evolve HG3, an artificial enzyme computationally designed to catalyze the not naturally occurring Kemp reaction.[121,122] As the initial activity of HG3 was far below natural enzymes, it was evolved over the course of a significant engineering campaign, totaling 17 rounds and introducing 17 new mutations, resulting in the Variant HG3.17 (k_{cat}/K_m = 230'000 ± 20'000 s^{-1} M^{-1}).[15]

We chose HG3 as a target for evolution because the catalyzed reaction can be easily screened using a chromophore-based read-out. In addition, data on the original evolutionary trajectory toward HG3.17 is available, allowing for a direct evaluation of the performance of our filter/oligo pipeline. HG3 libraries were designed based on the strategy outlined in **Article III**, by removing mutations predicted to be destabilizing and including mutations from the consensus approach described in **1.2.2.1.2**. We designed 1600-1800 variants at each round to cover most of the active site and any mutation below a ~0.5 REU ΔΔG cutoff. We split each round of evolution into two sections: 1) finding hits and 2) recombining hits. To find hits, we aimed to evaluate roughly 2000 variants, in line with coverage/sampling rates reported by Codexis.[104] We then built simple combinatorial libraries from the best-performing variants, including the wild-type. The reason to focus on simple recombination rather than more elaborate schemes is because of the limitations described in the section about **Article III**. Finding the most optimal solution in each round is much less critical than quickly identifying favorable mutations for recombination to attain fitness targets.[104]

Within five rounds of evolution, 16 new mutations were introduced into HG3, resulting in the final variant HG3.R5. This variant shows similar activities as HG3.17 under screening conditions (~458 FIOP for HG3.R5 and ~396 for HG3.17). Notably, even though both Kemp eliminase variants exhibit similar activities, they are quite different from each other. HG3.R5 differs from HG3 by 16 mutations yet only shares one mutation with HG3.17, which contains 17 mutations that differ from HG3. This mutation is K50Q, which was identified as a key mutation in HG3.17. Previous studies hypothesized K50Q to stabilize the negative charge developing during the transition state.[15,123]

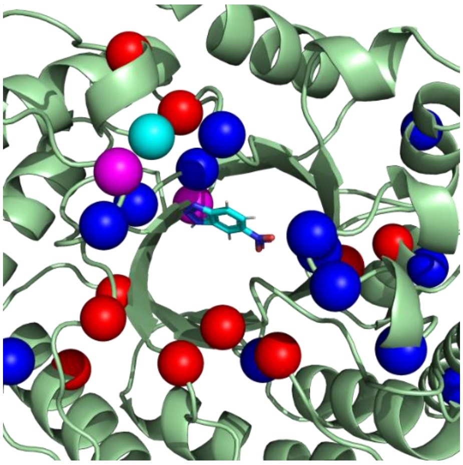


Figure 10: Comparison of the mutations of HG3.R5 and HG3.17. Mutations in the final HG3.R5 variant are colored in red. The mutations corresponding to HG3.17 are colored in blue. K50Q, the mutation that both variants have in common, is highlighted in turquoise. Pink indicates the sites that both variants substitute, though with different residues. Image from draft ***Article IV****.*

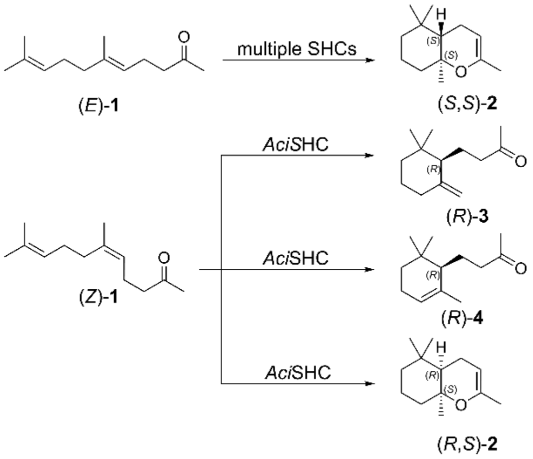
These results highlight the filter/oligo approach as an interesting protein engineering technique. Clearly, only a certain number of destabilizing mutations can be removed, leaving a relatively large sequence space to be screened. However, these mutations can be removed confidently. Additionally, oligo pools are a fascinating new technology that could completely change how protein engineers think about smart libraries. They are still far from perfect at the time of writing, suffering from high error rates, resulting in a large fraction of undesirable sequences.

Nevertheless, the pipeline described in **Article III** and applied in **Article IV** highlights the power of this technique. Rather than the 17 rounds described in the original publication,[15] the total amount of evolution rounds was reduced to 5, each requiring 8-10 weeks to complete. The final variant HG3.R5, even though distinct in its composition, exhibited the same catalytic activity as HG3.17 stimulating considerations about the underlying protein-fitness landscape and the structural factors governing catalysis.

2.3. Asymmetric cation-olefin monocyclization by engineered squalene-hopene cyclases (**Article V**)

Ionones are important components of the enticing scent of many flowers and fruits, such as violets, roses, and raspberries, and are commonly used in cosmetics and perfumes.[124,125] However, different isomeric forms can exhibit diverse olfactory profiles with varying odor thresholds.[126,127] As a result, only one isomer, or a defined mixture of them, is ideally employed in a given flavor and fragrance formulation, facilitating the need for selective and "natural" synthesis pathways. In **Article V**, we employed squalene-hopene cyclases (SHCs) to provide a novel approach for the highly enantioselective asymmetric synthesis of (*R*)-γ-dihydroionone (**3**) from the affordable industrial product geranylacetone. SHCs are capable of pre-folding linear terpenoids into specified chiral conformations, allowing for precise stereo control over polyene cyclizations,[128] and, just as importantly, have been shown to be highly evolvable.[129–132] This ease of reconfiguration makes them an ideal starting point to gain access to valuable monocyclic terpenoids starting from either (*E*/*Z*)-geranylacetone (**1**) or (*E*/*Z*)-pseudoionone.

Most research on SHCs focuses on a few highly studied variants, which were reported to convert our desired substrate into an undesired bicyclic product (**2**).[133,134] As such, we decided to expand SHC diversity by creating an exhaustive wild-type enzyme library of 31 SHCs evenly spread across all clades of the phylogenetic tree. Screening this library led to the identification of *Aci*SHC, which uniquely generated the monocyclic products γ-dihydroionone (**3**) and α-dihydroionone (**4**) with low conversions (0.7 % and 0.05 %) from (*E*/*Z*)-geranylacetone.



*Scheme 1: Transformations observed for AciSHC with (E/Z)-geranylacetone (**1**).*

As the product γ-dihydroionone, is a product of interest for the flavor and fragrance industry, we decided to engineer *Aci*SHC towards better activity and selectivity to convert **1** into **3**. We focused our efforts on the active site to improve substrate pre-folding and reduce space in the active pocket to limit unproductive binding modes. Thus, we selected 14 sites for full single-site saturation, leading to several mutants with improved conversion (between ~3 to 5.4 fold compared to the wild-type) at sites A169X, P263X, A310X, G606X, and I613X. We then selected to combine these beneficial mutations in a 5-site combinatorial library with a theoretical size of 288 variants. The combined mutations revealed multiple improved enzymes, with the best variant, called *Aci*SHC_R2.1 (A169P, A310M, G606C, I613V), improving upon the wild-type by more than 30-fold, achieving a conversion of **1** into **3** of 21.4 %. We then improved the yields further to 79 % through process optimization. Notably, product **3** was almost exclusively produced from (*Z*)-**1**. The best-performing engineered *Aci*SHC variants were able to differentiate between the geometric geranylacetone isomers, forming the monocyclic products (*R*)-**3** and (*R*)-**4** from (*Z*)-(**1**) and the bicyclic product from (*E*)-(**1**). Based on the obtained knowledge, our partners at Givaudan set out to access (*S*)-**3**, a valuable intermediate for alpha-ambrinol.

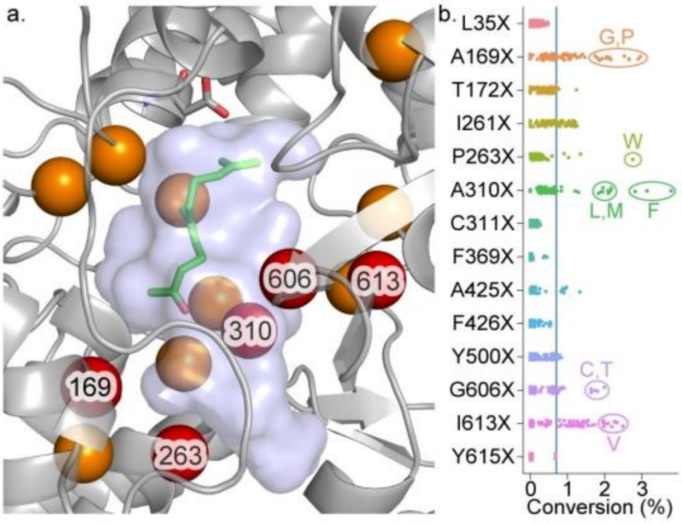


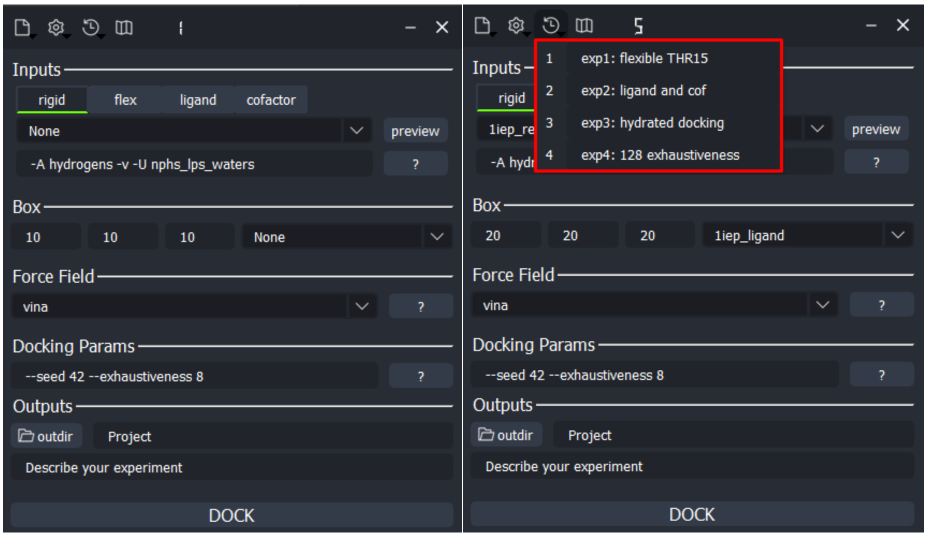
*Figure 11: a.) Homology model of AciSHC with the docked substrate (Z)-**1**. Sites around the active pocket, close to the substrate, and residues that might improve the productive binding of the smaller substrate were selected for single-site saturation mutagenesis (orange). The best-performing variants at sites A169X, P263X, A310X, G606X, and I613X were then further investigated in a combinatorial library. b.) Overview of the observed conversion for the different single-site saturation libraries. The blue horizontal line represents the wild-type activity. Image from Eichenberger et al. 2021.*[46]

This was accomplished with a masked (*E*)-**1** substrate to prevent the second cyclization step, resulting in the synthesis of (*S*)-**3** with perfect enantioselectivity. The study indicates that SHCs could provide a path toward enantioselective and stereodivergent transformations of geometric isomers. Additionally, SHCs can provide access to both enantiomers of a desired product, including the valuable building block (*S*)-**3,** through suitable substrate engineering and downstream processing.

2.4. Application note: AutoDock Vina plugin for PyMol (**Article VI**)

A major consideration with computational tools is their accessibility. One such example is molecular docking, a common denominator of many protein engineering projects and a valuable tool for studying the interactions between molecules (such as a receptor and a ligand).[135] As such, we developed AlphaDock, a PyMol plugin to access the powerful and widely popular tool AutoDock Vina,[29] which was recently expanded to include various new features. AlphaDock focuses on reducing user friction and application friction by simplifying the setup process and offloading computationally expensive work to more powerful workstation computers. Additionally, we place a big focus on reproducibility and traceability. All docking experiments are carefully logged, and a detailed history of all runs can be browsed and viewed. It is important to note that all required programs are containerized. This allows for an easy setup and ensures that results are always reproducible, irrespective of the computer the program is run.

The Plugin and more information are available at: https://github.com/ccbiozhaw/dock.



None

-A hydrogens -v -U nphs_lps_waters

10 10 10 None

vina

--seed 42 --exhaustiveness 8

Project

Describe your experiment

1 exp1: flexible THR15
2 exp2: ligand and cof
3 exp3: hydrated docking
4 exp4: 128 exhaustiveness

1iep_re

-A hydr

20 20 20 1iep_ligand

vina

--seed 42 --exhaustiveness 8

Project

Describe your experiment

Figure 12: AlphaDock Plugin for PyMol. All of the available options of AutoDock Vina 1.2+ can be accessed and altered. Each docking experiment is stored and can be accessed at any time through the history menu.

3. Summary

The field of protein engineering has experienced significant growth in recent decades, resulting in numerous successful applications of biocatalysts. Nevertheless, reconfiguring a protein for a specific task remains a complex and intricate endeavor. It is challenging to navigate the vast protein landscape, where peaks high enough to be relevant for an industrial application might be few and far between. In fact, the sheer magnitude of this space makes a comprehensive search entirely impossible. Ultimately, in protein engineering, the goal has to be attaining the highest peak given the available resources.

Article I explores successful applications of machine learning, a potentially powerful tool when aiming to reduce sequence space, in protein engineering. **Article II** builds on this exploratory work to configure the activity and regioselectivity of the halogenase WelO5*. Based on only a subset of all possible combinations, machine learning techniques were able to predict improved variants accurately.

Interestingly, some of the most prominent evolution campaigns (also described in **Article I**) considered the entire engineering process rather than just individual aspects. Ideally, we would design the perfect enzyme computationally, order a single gene and move to industrial production. However, for now, that is an utopic thought. As such, we must consider the practical limitations of protein engineering. How significant is the impact of a (computational) method? To make an enzyme industrially relevant, we might need to evolve it over multiple rounds. Can we justify the additional cost and delays introduced by sequencing and predicting? How can we combine these workflows with advances in automation? Do we need to find an "optimum" at every step, or should we move forward with a "good enough" variant? Such practical considerations lead to **Articles III** and **IV**. Predicting the variants that do not improve fitness rather than those that do, has a lot of beneficial attributes. Most importantly, it seems to be a much easier task. Additionally, it allows for significant generalization. Whether the objective is to improve activity, stability, or enantioselectivity, if a mutation causes the enzyme not to express or fold adequately, it is unlikely to improve function. Notably, recent advances in gene synthesis have made it possible to construct these libraries economically. This combination, removing destabilizing mutations and creating the remaining variants, could constitute a new, very efficient way of protein engineering. It also directly addresses most of the questions posed above. We can remove destabilizing predictions reliably. Furthermore, oligo pools are cheap and have short turnaround times, allowing us to conclude a round of evolution in ~8 weeks (which can easily be optimized), and their flexibility perfectly integrates with automation platforms. We can modulate the strictness of filtering to account for screening limitations without introducing additional burden in library construction.

We validated this strategy in **Article IV**, where we re-evolved the Kemp eliminase HG3. After five rounds of evolution, we arrived at the variant HG3.R5, which shows slightly improved activity (under screening conditions) than HG3.17. While not necessarily the focus of the study, HG3.R5 took a completely different evolutionary trajectory than HG3.17, and the two variants only have one mutation in common. This work demonstrates the potential of the introduced filtering strategy. Additional work will be required to optimize the process further, yet the results indicate a bright future for the methodology.

While developing new tools and strategies is essential, providing publicly accessible tools to distribute them should not be neglected. Without that, the reach of an idea will be severely limited, and exciting techniques might get lost. We explored this additional extension to method development in **Articles III and VI** with a website and a PyMol plugin.

The enzyme engineering workflows of the future will most likely combine various disciplines and techniques to achieve enhanced efficiency, robustness, and generalizability. Rather than improving individual aspects in a vacuum, it will be essential to consider the process as a whole and how different parts interact. In this spirit, the finding of this thesis indicates the potential of combining advances in gene synthesis with computational techniques to build novel enzyme engineering pipelines.

4. References

1. Schmid, A. *et al.* Industrial biocatalysis today and tomorrow. *Nature* **409**, 258–268 (2001).
2. Leisola, M. & Turunen, O. Protein engineering: Opportunities and challenges. *Appl Microbiol Biotechnol* **75**, 1225–1232 (2007).
3. Grunwald, P. Biocatalysis: Biochemical fundamentals and applications. *Imp Col Press,* 1–10 (2009).
4. Magliery, T. J. Protein stability: Computation, sequence statistics, and new experimental methods. *Curr Opin Struct Biol* **33**, 161–168 (2015).
5. Lutz, S. Beyond directed evolution-semi-rational protein engineering and design. *Curr Opin Biotechnol* **21**, 734–743 (2010).
6. Reetz, M.T. A Method for Rapid Directed Evolution. In Protein engineering handbook. Edited by S. Lutz and U. T. Bornscheuer. *Wiley*, *409-439* (2008).
7. da Silva Amatto, I. V. *et al.* Enzyme engineering and its industrial applications. *Biotechnol Appl Biochem* **69**, 389–409 (2022).
8. Wang, Y. *et al.* Directed evolution: Methodologies and applications. *Chem Rev* **121**, 12384–12444 (2021).
9. The Nobel Prize in Chemistry 2018. https://www.nobelprize.org/prizes/chemistry/2018/summary/ (2018).
10. Park, S. *et al.* Focusing mutations into the *P. fluorescens* esterase binding site increases enantioselectivity more effectively than distant mutations. *Chem Biol* **12**, 45–54 (2005).
11. Morley, K. & Kazlauskas, R. Improving enzyme properties: When are closer mutations better? *Trends Biotechnol* **23**, 231–237 (2005).
12. Steiner, K. & Schwab, H. Recent advances in rational approaches for enzyme engineering. *Comput Struct Biotechnol J* **2**, e201209010 (2012).
13. Arnold, F. H. Innovation by evolution: Bringing new chemistry to life (Nobel lecture). *Angew Chem Int Ed* **58**, 14420–14426 (2019).
14. Bloom, J. D., Labthavikul, S. T., Otey, C. R. & Arnold, F. H. Protein stability promotes evolvability. *Proc Natl Acad Sci U S A* **103**, 5869–5874 (2006).
15. Blomberg, R. *et al.* Precision is essential for efficient catalysis in an evolved Kemp eliminase. *Nature* **503**, 418–421 (2013).
16. Patsch, D. & Buller, R. Improving enzyme fitness with machine Learning. *Chimia* **77**, 116 (2023).
17. Reetz, M. Making enzymes suitable for organic chemistry by rational protein design. *ChemBioChem* **23**, e202200049 (2022).
18. Kazlauskas, R. & Bornscheuer, U. Finding better protein engineering strategies. *Nat Chem Biol* **5**, 526–529 (2009).

19. Janda, J. O., Busch, M., Kück, F., Porfenenko, M. & Merkl, R. CLIPS-1D: Analysis of multiple sequence alignments to deduce for residue-positions a role in catalysis, ligand-binding, or protein structure. *BMC Bioinform* **13**, 55 (2012).

20. Sternke, M., Tripp, K. W. & Barrick, D. Consensus sequence design as a general strategy to create hyperstable, biologically active proteins. *Proc Natl Acad Sci U S A* **166**, 11275–11284 (2019).

21. Jacak, R., Leaver-Fay, A. & Kuhlman, B. Computational protein design with explicit consideration of surface hydrophobic patches. *Proteins* **80**, 825–838 (2012).

22. Borgo, B. & Havranek, J. J. Automated selection of stabilizing mutations in designed and natural proteins. *Proc Natl Acad Sci U S A* **109**, 1494–1499 (2012).

23. Childers, M. & Daggett, V. Insights from molecular dynamics simulations for protein design. *Mol Syst Des Eng* **2**, 9-33 (2017).

24. Goldenzweig, A. *et al.* Automated structure- and sequence-based design of proteins for high bacterial expression and stability. *Mol Cell* **63**, 337–346 (2016).

25. Korendovych, I. Rational and semirational protein design. *Methods mol biol* **1685,** 15–23 (2018).

26. Dvorak, P. Methods of directed evolution and their application for engineering of haloalkane dehalogenases. (2007). doi:10.13140/2.1.3648.0006

27. Bornscheuer, U. T. *et al.* Engineering the third wave of biocatalysis. *Nature* **485**, 185–194 (2012).

28. Denard, C. A., Ren, H. & Zhao, H. Improving and repurposing biocatalysts via directed evolution. *Curr Opin Chem Biol* **25**, 55–64 (2015).

29. Eberhardt, J., Santos-Martins, D., Tillack, A. F. & Forli, S. AutoDock Vina 1.2.0: New docking methods, expanded force field, and python bindings. *J Chem Inf Model* **61**, 3891–3898 (2021).

30. Salentin, S., Schreiber, S., Haupt, V. J., Adasme, M. F. & Schroeder, M. PLIP: Fully automated protein-ligand interaction profiler. *Nucleic Acids Res* **43**, 443–447 (2015).

31. Zhang, Z., Li, Y., Lin, B., Schroeder, M. & Huang, B. Identification of cavities on protein surface using multiple computational approaches for drug binding site prediction. *Bioinformatics* **27**, 2083–2088 (2011).

32. Chovancova, E. *et al.* CAVER 3.0: A tool for the analysis of transport pathways in dynamic protein structures. *PLoS Comput Biol* **8**, e1002708 (2012).

33. Brezovsky, J. *et al.* Software tools for identification, visualization and analysis of protein tunnels and channels. *Biotechnol Adv* **31**, 38–49 (2013).

34. Sebestova, E., Bendl, J., Brezovsky, J. & Damborský, J. Computational tools for designing smart libraries. *Methods Mol Biol* **1179**, 291–314 (2014).

35. Bommarius, A. S. & Paye, M. F. Stabilizing biocatalysts. *Chem Soc Rev* **42**, 6534–6565 (2013).

36. Lehmann, M. & Wyss, M. Engineering proteins for thermostability: The use of sequence alignments versus rational design and directed evolution. *Curr Opin Biotechnol* **12**, 371–375 (2001).

37. Yu, H. & Huang, H. Engineering proteins for thermostability through rigidifying flexible sites (RFS). *Biotechnol Adv* **32**, 308-315 (2013).

38. Wijma, H., Floor, R. & Janssen, D. Structure- and sequence-analysis inspired engineering of proteins for enhanced thermostability. *Curr Opin Struct Biol* **23**, 588-954 (2013).

39. Luo, Y. *et al.* ECNet is an evolutionary context-integrated deep learning framework for protein engineering. *Nat Commun* **12**, 5743 (2021).

40. Matsuno, K. *et al.* Identification of a new series of STAT3 inhibitors by virtual screening. *ACS Med Chem Lett* **1**, 371–375 (2010).

41. Kiss, R. *et al.* Discovery of novel human histamine H4 receptor ligands by large-scale structure-based virtual screening. *J Med Chem* **51**, 3145–3153 (2008).

42. Ren, J. X. *et al.* Discovery of novel Pim-1 kinase inhibitors by a hierarchical multistage virtual screening approach based on svm model, pharmacophore, and molecular docking. *J Chem Inf Model* **51**, 1364–1375 (2011).

43. Clyde, A. *et al.* High-throughput virtual screening and validation of a SARS-CoV-2 main protease noncovalent inhibitor. *J Chem Inf Model* **62**, 116–128 (2022).

44. Naik, B. *et al.* High throughput virtual screening reveals SARS-CoV-2 multi-target binding natural compounds to lead instant therapy for COVID-19 treatment. *Int J Biol Macromol* **160**, 1-17 (2020).

45. Reetz, M. T., Wang, L. W. & Bocola, M. Directed evolution of enantioselective enzymes: Iterative cycles of CASTing for probing protein-sequence space. *Angew Chem Int Ed* **45**, 1236–1241 (2006).

46. Eichenberger, M. *et al.* Asymmetric cation–Olefin monocyclization by engineered squalene–hopene cyclases. *Angew Chem Int Ed* **60**, 26080-26086 (2021).

47. Büchler, J. *et al.* Algorithm-aided engineering of aliphatic halogenase WelO5* for the asymmetric late-stage functionalization of soraphens. *Nat Commun* **13**, 371 (2022).

48. Pagadala, N. S., Syed, K. & Tuszynski, J. Software for molecular docking: a review. *Biophys Rev* **9**, 91–102 (2017).

49. Trott, O. & Olson, A. Software news and update AutoDock Vina: Improving the speed and accuracy of docking with a new scoring function, efficient optimization, and multithreading. *J Comput Chem* **31**, 455–461 (2010).

50. AutoDock Vina program. *https://github.com/ccsb-scripps/AutoDock-Vina (2023)*.

51. Porebski, B. & Buckle, A. Consensus protein design. *Protein Eng Des Sel* **29**, 245-51 (2016).

52. Bendl, J. *et al.* HotSpot Wizard 2.0: automated design of site-specific mutations and smart libraries in protein engineering. *Nucleic Acids Res* **44**, 479–487 (2016).

53. Sumbalova, L., Stourac, J., Martinek, T., Bednar, D. & Damborsky, J. HotSpot Wizard 3.0: Web server for automated design of mutations and smart libraries based on sequence input information. *Nucleic Acids Res* **46**, 356–362 (2018).

54. Altschul, S. F., Gertz, E. M., Agarwala, R., Schäffer, A. A. & Yu, Y. K. PSI-BLAST pseudocounts and the minimum description length principle. *Nucleic Acids Res* **37**, 815–824 (2009).

55. Finn, R. D. *et al.* The Pfam protein families database: Towards a more sustainable future. *Nucleic Acids Res* **44**, 279–285 (2016).

56. Wilson, D. *et al.* SUPERFAMILY - Sophisticated comparative genomics, data mining, visualization and phylogeny. *Nucleic Acids Res* **37**, 380-386 (2009).

57. Apweiler, R. The universal protein resource (UniProt). *Nucleic Acids Res* **36**, 190-195 (2008).

58. Acland, A. *et al.* Database resources of the national center for biotechnology information. *Nucleic Acids Res* **42**, 8-20 (2014).

59. Cerdobbel, A. *et al.* Increasing the thermostability of sucrose phosphorylase by a combination of sequence- and structure-based mutagenesis. *Protein Eng Des Sel* **24**, 829–834 (2011).

60. Jochens, H., Aerts, D. & Bornscheuer, U. T. Thermostabilization of an esterase by alignment-guided focussed directed evolution. *Protein Eng Des Sel* **23**, 903–909 (2010).

61. Qu, G., Sun, Z. & Reetz, M. T. Iterative saturation mutagenesis for semi-rational enzyme design. In Protein Engineering*. Wiley,* 105–132 (2021).

62. Sora, V. *et al.* RosettaDDGPrediction for high-throughput mutational scans: From stability to binding. *Protein Sci* **32**, e4527 (2022).

63. Kellogg, E. H., Leaver-Fay, A. & Baker, D. Role of conformational sampling in computing mutation-induced changes in protein structure and stability. *Proteins* **79**, 830–838 (2011).

64. Frenz, B. *et al.* Prediction of protein mutational free energy: Benchmark and sampling Improvements increase classification accuracy. *Front Bioeng Biotechnol* **8**, 558247 (2020).

65. Allen, M. & Tildesley, D. Computer simulation of liquids. *Oxford*, 1-45 (2017).

66. Wildman, J., Repiščák, P., Paterson, M. J. & Galbraith, I. General force-field parametrization scheme for molecular dynamics simulations of conjugated materials in solution. *J Chem Theory Comput* **12**, 3813–3824 (2016).

67. Teague, S.J. Implications of protein flexibility for drug discovery. *Nat Rev Drug Discov* **2**, 527–541 (2003).

68. Nabuurs, S. B., Wagener, M. & de Vlieg, J. A flexible approach to induced fit docking. *J Med Chem* **50**, 6507–6518 (2007).

69. Allegra, M. *et al.* Evaluation of the IKKβ Binding of Indicaxanthin by induced-fit docking, binding pose metadynamics, and molecular dynamics. *Front Pharmacol* **12**, 701568 (2021).

70. Guterres, H. & Im, W. Improving protein-ligand docking results with high-throughput molecular dynamics simulations. *J Chem Inf Model* **60**, 2189–2198 (2020).

71. Khan, S. H. *et al.* Protein folding: Molecular dynamics simulations and *in vitro* studies for probing mechanism of urea- and guanidinium chloride-induced unfolding of horse cytochrome-c. *Int J Biol Macromol* **122**, 695–704 (2019).

72. Sun, Z., Liu, Q., Qu, G., Feng, Y. & Reetz, M. Utility of B-factors in protein science: interpreting rigidity, flexibility, and internal motion and engineering thermostability. *Chem Rev* **119**, 1626-1665 (2019).

73. Park, H. J., Park, K., Kim, Y. H. & Yoo, Y. J. Computational approach for designing thermostable *Candida antarctica* lipase B by molecular dynamics simulation. *J Biotechnol* **192**, 66–70 (2014).

74. Yokota, A., Takahashi, H., Takenawa, T. & Arai, M. Probing the roles of conserved arginine-44 of *Escherichia coli* dihydrofolate reductase in its function and stability by systematic sequence perturbation analysis. *Biochem Biophys Res Commun* **391**, 1703–1707 (2010).

75. Fredricksen, R. S. & Swenson, C. Relationship between stability and function for isolated domains of troponin C. *Biochemistry* **35**, 14012–14026 (1996).

76. Jomain, J.-B. *et al.* Structural and thermodynamic bases for the design of pure prolactin receptor antagonists: X-ray structure of Del1-9-G129R-hPRL. *J Biol Chem* **282**, 33118–33131 (2007).

77. Torrado, M. *et al.* Role of conserved salt bridges in homeodomain stability and DNA binding. *J Biol Chem* **284**, 23765–23779 (2009).

78. Zakrzewska, M., Krowarsch, D., Wiedlocha, A., Olsnes, S. & Otlewski, J. Highly stable mutants of human fibroblast growth factor-1 exhibit prolonged biological action. *J Mol Biol* **352**, 860–875 (2005).

79. Kragelund, B. B. *et al.* Hydrophobic core substitutions in calbindin d 9k : effects on ca 2+ binding and dissociation. *Biochemistry* **37**, 8926–8937 (1998).

80. Julenius, K., Thulin, E., Linse, S. & Finn, B. Hydrophobic core substitutions in calbindin d 9k : effects on stability and structure. *Biochemistry* **37**, 8915–8925 (1998).

81. Seelig, B. & Szostak, J. Selection and evolution of enzymes from a partially randomized non-catalytic scaffold. *Nature* **448**, 828–831 (2007).

82. Kiss, G., Çelebi-Ölçüm, N., Moretti, R., Baker, D. & Houk, K. Computational enzyme design. *Angew Chem Int Ed* **52**, 5700-5725 (2013).

83. Porebski, B. T. & Buckle, A. M. Consensus protein design. *Protein Eng Des Sel* **29**, 245–251 (2016).

84. Kaushik, M. *et al.* Protein engineering and *de novo* designing of a biocatalyst. *J Mol Recognit* **29**, 499– 503 (2016).

85. Li, D., Wu, Q. & Reetz, M. Focused rational iterative site-specific mutagenesis (FRISM). *Meth Enzymol* vol. **643** 225-242 (2020).

86. Reetz, M. Laboratory evolution of stereoselective enzymes: a prolific source of catalysts for asymmetric reactions. *Angew Chem Int Ed Engl* **50**, 138–174 (2011).

87. Reetz, M. Biocatalysis in organic chemistry and biotechnology: past, present, and future. *J Am Chem Soc* **135**, 12480-12496 (2013).

88. Balke, K., Beier, A. & Bornscheuer, U. Hot spots for the protein engineering of Baeyer-Villiger monooxygenases. *Biotechnol Adv* **36**, 247-263 (2017).

89. Acevedo-Rocha, C., Ferla, M. & Reetz, M. Directed evolution of proteins based on mutational scanning. *Methods mol biol* **1685,** 87–128 (2018).

90. Reetz, M. T., Kahakeaw, D. & Lohmer, R. Addressing the numbers problem in directed evolution. *ChemBioChem* **9**, 1797–1804 (2008).

91. Wittmann, B., Johnston, K., Wu, Z. & Arnold, F. Advances in machine learning for directed evolution. *Curr Opin Struct Biol* **69**, 11–18 (2021).

92. Romero, P. A., Krause, A. & Arnold, F. H. Navigating the protein fitness landscape with Gaussian processes. *Proc Natl Acad Sci U S A* **110**, 193-201 (2013).

93. Cadet, F. *et al.* A machine learning approach for reliable prediction of amino acid interactions and its application in the directed evolution of enantioselective enzymes. *Sci Rep* **8**, 16757 (2018).

94. Li, Y. *et al.* A diverse family of thermostable cytochrome P450s created by recombination of stabilizing fragments. *Nat Biotechnol* **25**, 1051–1056 (2007).

95. Klesmith, J., Bacik, J.-P., Wrenbeck, E., Michalczyk, R. & Whitehead, T. Trade-offs between enzyme fitness and solubility illuminated by deep mutational scanning. *Proc Natl Acad Sci U S A* **114**, 201614437 (2017).

96. Singh, S. P. *et al.* Machine translation using deep learning: An overview. *Int Conf Electr Eng Inform Commun Technol,* 162–167 (2017).

97. Furtado, F. & Singh, A. Movie recommendation system using machine learning. *International Journal of Research in Industrial Engineering* **9**, 84–98 (2020).

98. Kryshtafovych, A., Schwede, T., Topf, M., Fidelis, K. & Moult, J. Critical assessment of methods of protein structure prediction (CASP)—Round XIV. *Proteins* **89**, 1607–1617 (2021).

99. Jumper, J. *et al.* Highly accurate protein structure prediction with AlphaFold. *Nature* **596**, 583–589 (2021).

100. Anfinsen, C. Principles that govern the folding of protein chains. *Science* **181**, 223–230 (1973).

101. Lin, Z. *et al.* Evolutionary-scale prediction of atomic-level protein structure with a language model. *Science* **379**, 1123–1130 (2023).

102. Yang, K., Wu, Z. & Arnold, F. Machine-learning-guided directed evolution for protein engineering. *Nat Methods* **16**, 1 (2019).

103. Feng, X., Sanchis, J., Reetz, M. & Rabitz, H. Enhancing the efficiency of directed evolution in focused enzyme libraries by the adaptive substituent reordering algorithm. *Chem. Eur. J.* **18**, 5646–5654 (2012).

104. Alvizo, O. *et al.* Directed evolution of an ultrastable carbonic anhydrase for highly efficient carbon capture from flue gas. *Proc Natl Acad Sci U S A* **111**, 16436–16441 (2014).

105. Sanders, H. & Saxe, J. Garbage in, garbage out: how purportedly great ml models can be screwed up by bad data. (2017).

106. Kawashima, S. *et al.* AAindex: Amino acid index database, progress report 2008. *Nucleic Acids Res* **36**, 202-205 (2008).

107. Blackledge, J.: Digital signal processing (second edition). *Horwood Publishing*, vol: ISBN: 1-904275-26-5. (2006).

108. Rives, A. et al. Biological structure and function emerge from scaling unsupervised learning to 250 million protein sequences. *Proc Natl Acad Sci U S A* **118**, e2016239118 (2021).

109. Fox, R. J. *et al.* Improving catalytic function by ProSAR-driven enzyme evolution. *Nat Biotechnol* **25**, 338–344 (2007).

110. Shroff, R. *et al.* Discovery of novel gain-of-function mutations guided by structure-based deep learning. *ACS Synth Biol* **9**, 2927–2935 (2020).

111. Hayashi, T. *et al.* Evolved aliphatic halogenases enable regiocomplementary C−H functionalization of a pharmaceutically relevant compound. *Adv Mater* **131**, 18706–18711 (2019).

112. Saito, Y. *et al.* Machine-learning-guided mutagenesis for directed evolution of fluorescent proteins. *ACS Synth Biol* **7**, 2014–2022 (2018).

113. Ma, E. J. *et al.* Machine-directed evolution of an imine reductase for activity and stereoselectivity. *ACS Catal* **11**, 12433–12445 (2021).

114. Kuiper, B. P., Prins, R. C. & Billerbeck, S. Oligo pools as an affordable source of synthetic DNA for cost-effective library construction in protein- and metabolic pathway engineering. *ChemBioChem* **23,** e202100507 (2022).

115. Serrano-Vega, M., Magnani, F., Shibata, Y. & Tate, C. Conformational thermostabilization of the 1-adrenergic receptor in a detergent-resistant form. *Proc Natl Acad Sci U S A* **105**, 877–882 (2008).

116. Magnani, F., Shibata, Y., Serrano-Vega, M. & Tate, C. Co-evolving stability and conformational homogeneity of the human adenosine A2A receptor. *Proc Natl Acad Sci U S A* **105**, 10744–10749 (2008).

117. Novick, S. J. *et al.* Engineering an amine transaminase for the efficient production of a chiral sacubitril precursor. *ACS Catal* **11**, 3762–3770 (2021).

118. Faber, M. S. *et al.* Saturation mutagenesis genome engineering of infective φx174 bacteriophage via unamplified oligo pools and golden gate assembly. *ACS Synth Biol* **9**, 125–131 (2020).

119. Kitzman, J. O., Starita, L. M., Lo, R. S., Fields, S. & Shendure, J. Massively parallel single-amino-acid mutagenesis. *Nat Methods* **12**, 203–206 (2015).

120. Steiner, P., Baumer, Z. & Whitehead, T. A Method for user-defined mutagenesis by integrating oligo pool synthesis technology with nicking mutagenesis. *Bio Protoc* **10**, e3697 (2020).

121. Röthlisberger, D. *et al.* Kemp elimination catalysts by computational enzyme design. *Nature* **453**, 190–195 (2008).

122. Privett, H. K. *et al.* Iterative approach to computational enzyme design. *Proc Natl Acad Sci U S A* **109**, 3790–3795 (2012).

123. Broom, A. *et al.* Ensemble-based enzyme design can recapitulate the effects of laboratory directed evolution in silico. *Nat Commun* **11**, 4808 (2020).

124. Zhang, C. Biosynthesis of carotenoids and apocarotenoids by microorganisms and their industrial potential. In Progress in Carotenoid Research. *IntechOpen,* 85–105 (2018).

125. Petroni, K. & Tonelli, C. Recent advances on the regulation of anthocyanin synthesis in reproductive organs. *Plant Sci* **181**, 219–229 (2011).

126. Fuganti, C., Serra, S. & Zenoni, A. Synthesis and olfactory evaluation of (+)- and (-)-γ-ionone. *Helv Chim Acta* **83**, 2761–2768 (2000).

127. Brenna, E., Fuganti, C., Serra, S. & Kraft, P. Optically active ionones and derivatives: preparation and olfactory properties. *Eur J Org Chem* **2002**, 967–978 (2010).

128. Siedenburg, G. & Jendrossek, D. Squalene-hopene cyclases. *Appl Environ Microbiol* **77**, 3905–3915 (2011).

129. Hammer, S. C., Syrén, P. O., Seitz, M., Nestl, B. M. & Hauer, B. Squalene hopene cyclases: Highly promiscuous and evolvable catalysts for stereoselective CC and CX bond formation. *Curr Opin Chem Biol* **17**, 293–300 (2013).

130. Bastian, S. A., Hammer, S. C., Kreß, N., Nestl, B. M. & Hauer, B. selectivity in the cyclization of citronellal introduced by squalene hopene cyclase variants. *ChemCatChem* **9**, 4364–4368 (2017).

131. Hammer, S. C., Syrén, P. O. & Hauer, B. Substrate pre-folding and water molecule organization matters for terpene cyclase catalyzed conversion of unnatural substrates. *ChemistrySelect* **1**, 3589–3593 (2016).

132. Hammer, S. C., Marjanovic, A., Dominicus, J. M., Nestl, B. M. & Hauer, B. Squalene hopene cyclases are protonases for stereoselective Brønsted acid catalysis. *Nat Chem Biol* **11**, 121–126 (2015).

133. Abe, I. Enzymatic synthesis of cyclic triterpenes. *Nat Prod Rep* **24**, 1311–1331 (2007).

134. Syrén, P. O., Henche, S., Eichler, A., Nestl, B. M. & Hauer, B. Squalene-hopene cyclases - evolution, dynamics and catalytic scope. *Curr Opin Struct Biol* **41**, 73–82 (2016).

135. Grosdidier, A., Zoete, V. & Michielin, O. SwissDock, a protein-small molecule docking web service based on EADock DSS. *Nucleic Acids Res* **39**, 270-277 (2011).

5. Author Contributions

Article I **Improving enzyme fitness with machine learning**

D. Patsch, R. Buller, *CHIMIA* **2023**, 77(3), 116. DOI: 10.2533/chimia.2023.116

D.P. and R.B. carried out the manuscript's research, conceptualization, and writing. R.B. initiated and supervised the project.

Article II **Algorithm-aided engineering of aliphatic halogenase WelO5* for the asymmetric late-stage functionalization of soraphens**

J. Büchler, S. Honda Malca, D. Patsch, M. Voss, N. J. Turner, U. T. Bornscheuer, O. Allemann, C. Le Chapelain, A. Lumbroso, O. Loiseleur, R, Buller, *Nat. Commun.,* **2022,** 13(1), 371. DOI: 10.1038/s41467-022-27999-1

J.B., A.L., C.L.C., O.L. and R.B. designed the research. J.B. carried out most of the experiments and performed the docking analysis. S.H.M. and M.V. constructed the combinatorial enzyme library. D.P. carried out the machine learning predictions. O.A. synthesized the soraphen analogs. C.L.C. analyzed the NMR structures. J.B., N.J.T., U.T.B., C.L.C., O.L. and R.B. discussed the results and wrote the manuscript.

Article III **LibGENiE – A bioinformatic pipeline for the design of information-enriched enzyme libraries**

D. Patsch, M. Eichenberger, M. Voss, U. T. Bornscheuer, R. Buller, submitted to *Comput. Struct. Biotechnol. J.*, **2023**.

D.P. and M.E. conceptualized the methodology. D.P. designed the research, collected data, implemented the software, analyzed the results, and wrote the manuscript. M.E., M.V., U.T.B. and R.B. helped write, discuss, and correct the manuscript. R.B. supervised the project.

Article IV **Efficient evolution of a Kemp eliminase**

D. Patsch, M. Voss, T. Schwander, U. T. Bornscheuer, R. Buller, article in preparation, **2023**.

D.P., M.V. and R.B. designed the research and approach. D.P. performed the computational aspects of the project. M.V. with help of D.P. carried out most of the experiments. D.P., T.S., U.T.B. and R.B. discussed the results and wrote the manuscript.

Article V **Asymmetric cation-olefin monocyclization by engineered squalene–hopene cyclases**

M. Eichenberger*, S. Hüppi*, D. Patsch*, N. Aeberl, R. Berweger, S. Dossenbach, E. Eichhorn, F. Flachsmann, L. Hortencio, F. Voirol, S. Vollenweider 5, U. T. Bornscheuer, R. Buller, *Angew. Chem. Int. Ed.*, **2021**, 60(50), 26080–6. DOI: 10.1002/anie.202108037

* equal contribution

M.E., S.H., D.P., E.E., F. F., and R.B. designed the research. M.E., S.H. and D.P. carried out most of the experiments and analyzed the docking. M.E., S.H., D.P. performed the library construction, screening, and evolution of *Aci*SHC variants, while N.A., R.B., S.D., E.E., F.F., L.H., F.V., and S.V. worked on the enantio-complementary product. All parties discussed the results and wrote the manuscript.

Article VI **AlphaDock – docking plugin**

D. Patsch, R. Buller, article in preparation, **2023**.

D.P., implemented the software, and wrote the manuscript. R.B. supervised the project.

David Patsch

Prof. Dr. Uwe T. Bornscheuer

Articles

Article I

doi:10.2533/chimia.2023.116 Chimia 77 (2023) 116–121 © D. Patsch, R. Buller

Improving Enzyme Fitness with Machine Learning

David Patsch[ab] and Rebecca Buller[a*]

Abstract: The combinatorial composition of proteins has triggered the application of machine learning in enzyme engineering. By predicting how protein sequence encodes function, researchers aim to leverage machine learning models to select a reduced number of optimized sequences for laboratory measurement with the aim to lower costs and shorten timelines of enzyme engineering campaigns. In this review, we highlight successful algorithm-aided protein engineering examples, including work carried out within NCCR Catalysis. In this context, we will discuss the underlying computational methods developed to improve enzyme properties such as enantioselectivity, regioselectivity, activity, and stability. Considering the rapid maturing of computational techniques, we expect that their continued application in enzyme engineering campaigns will be key to deliver additional powerful biocatalysts for sustainable chemical synthesis.

Keywords: Bioinformatics · Enzyme engineering · Halogenase · Industrial biocatalysis · Machine learning

David Patsch studied biology and received his BSc from the University of Innsbruck. He obtained his MSc degree in biotechnology from the Management Center Innsbruck. Since 2019 he is pursuing his PhD in the group of Rebecca Buller at the ZHAW.

Rebecca Buller is a biological chemist and Professor for Biotechnological Methods, Systems and Processes at the Zurich University of Applied Sciences. Rebecca Buller studied chemistry at the Westfälische – Wilhelms Universität Münster (D) and the University of California Santa Barbara (US). After completing her PhD with a focus on enzyme engineering at ETH Zurich (CH), Rebecca Buller accepted a position as laboratory head at the flavour and fragrance company Firmenich (CH). In 2015, she relocated to the Zurich University of Applied Sciences where she founded the Competence Center for Biocatalysis (CCBIO). Research in Rebecca Buller's laboratory focusses on the expansion of the biocatalytic toolbox by sourcing and engineering enzymes for synthetic applications.

1. Introduction

In optimal settings, enzymes can facilitate complex reactions with extraordinary specificity and selectivity.[1,2] However, practical reality usually differs from this ideal as wildtype enzymes are often just marginally stable in the selected reaction conditions[3] and perform at scales well below what is required to drive an industrial process. However, as enzymes are combinatorically composed from a limited set of simple building blocks, improved catalysts can be constructed in the laboratory by applying enzyme engineering strategies, among them the directed evolution of proteins. Consequently, engineered enzymes are harnessed in many industrial fields ranging from the fine chemical to the pharmaceutical sectors.[4–6]

Over the last decades, the technique of directed evolution has developed into a powerful tool (Nobel prize for chemistry 2018)[7] and today, it is routinely applied to tailor critical protein properties.[4,8] Directed evolution mimics nature's selection process in the laboratory through iterative cycles of gene diversification and selection of the encoded protein variants generating enzyme lineages with new or improved functions.[9] However, unlike nature, which selects for survival or reproduction, directed evolution can be used to precisely tailor desired protein traits.[10] In this context, astounding improvements in target biological functions for several different enzyme families have been achieved, including activity,[11–13] stereoselectivity,[14,15] thermostability,[16] and solvent tolerance.[17] Strikingly, these studies screened only a relatively small fraction of the target protein's underlying sequence space, raising the question of whether better sequence solutions would, in principle, exist for the function of interest. Unfortunately, such a question cannot easily be answered experimentally: Full randomization of a small protein consisting, for example, of 100 amino acids leads to a search space of sequences that is larger than the estimated number of atoms in the universe.[18] Even the targeted randomization of predefined positions within a protein quickly leads to a screening bottleneck: While replacing a single amino acid position with all other natural amino acids yields an experimentally manageable library size of 20^1 variants, combinatorically investigating as little as five sites in a protein already leads to a library size of 20^5. Clearly, it is difficult to experimentally screen such large libraries exhaustively, even when using advanced automation. In addition, most mutations introduced into a protein are either neutral or unfavorable,[19] leading to an even more inefficient sampling of the sequence space. To address the numbers problem in protein engineering, researchers are increasingly interested in implementing computational techniques, such as molecular dynamics simulations,[20] phylogeny, docking,[21,22] and, more recently, machine learning (ML) (Fig. 1).[23]

ML, in particular, has emerged as a powerful and versatile tool for various applications, many of which affect our daily lives, such as translating languages[24] or recommending what movies

---

Correspondence: Prof. R. Buller[a], E-Mail: rebecca.buller@zhaw.ch
[a]Institute of Chemistry and Biotechnology, Zurich University of Applied Sciences, CH-8820 Wädenswil, Switzerland; [b]Institute of Biochemistry, Dept. of Biotechnology & Enzyme Catalysis, Greifswald University, Felix-Hausdorff-Strasse 4, D17487 Greifswald, Germany

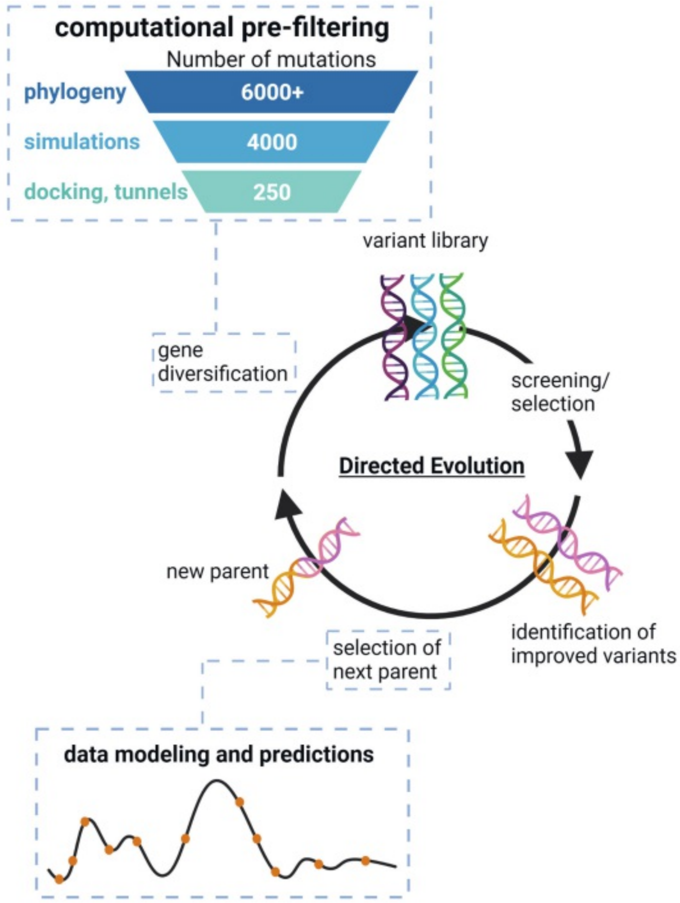


Fig. 1. Integration of *in silico* tools into directed evolution of proteins. As the random generation of genetic diversity is often inefficient when targeting to improve a desired function, information from various bioinformatic sources, such as phylogeny, docking, tunnels, and ML tools, can be used to build 'smart' enzyme libraries. Additionally, ML methods might be able to learn the underlying enzyme fitness landscape and suggest improved variants which have not yet been experimentally screened. Image created with BioRender.com.

to watch next.[25] Looking forward, ML is expected to profoundly impact the field of protein engineering as well. In contrast to traditional directed evolution, which discards information except if related to the most beneficial mutations, ML techniques can rely on all generated data to speed up the evolution process. This acceleration might be achieved by learning a function representing the underlying protein landscape from a set of sequence-fitness pairs. Based on this function, additional variants can be evaluated computationally, allowing exploration of the sequence space at a scale that cannot be achieved through laboratory experiments alone.[26] The potential benefits of ML make it an attractive research objective, and multiple attempts to apply it to protein engineering have been made. This report is by no means meant to cover them exhaustively but instead focuses on work related to research carried out in the frame of NCCR Catalysis.

2. ML-aided Optimization of Enzyme Stereoselectivity

From an organic chemist's perspective, facilitating the tailoring of the stereo- and regioselectivity of enzymes might be one of the most exciting applications of ML in protein engineering.[27] In this context, a first ML-driven study to improve enantioselectivity for the selective ring opening of a racemic mixture of glycidyl phenyl ether catalyzed by an epoxide hydrolase from *Aspergillus niger* (ANEH) was published in 2012.[21,28] More selective ANEH variants were predicted through the adaptive substituent reordering algorithm (ASRA),[29,30] which focuses on finding beneficial regions in a combinatorial enzyme library with minimal screening effort. The underlying principle of the approach is to first evaluate a small subset of all possible variants of a combinatorial enzyme library experimentally before reordering the amino-acid pairs to maximize the smoothness of the measured property landscape (Fig. 2).[28] Unlike the traditional quantitative structure–activity relationships (QSAR), ASRA does not make explicit assumptions about linearity, additivity, or specific relationships between structure and function. It only relies on the hypothesis that the underlying protein landscape is, to some extent, smooth.[31,32] This is an assumption ASRA shares with most, if not all, computational approaches and, from our experience, represents a valid bias in protein engineering in most cases. Within the ANEH study,[28] ASRA was shown to be a powerful tool for obtaining reliable estimates about areas of interest within the sizable sequence space that arises from evaluating variants combinatorically. Notably, ASRA did not rely on complex protein/residue descriptors making the algorithm a compelling starting point for protein engineering campaigns.

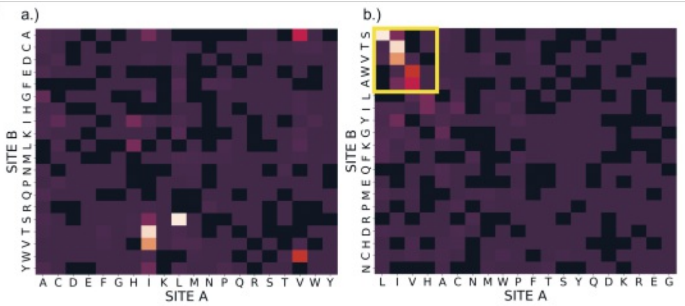


Fig. 2. Application of the ASRA algorithm: First, a random subset of a two-site combinatorial library is screened (left). Next, the amino acid pairs are rearranged to maximize the smoothness of the fitness landscape (right). This rearrangement highlights beneficial regions (yellow box in right plot) to explore in a library of reduced size. In this representation, black squares denote amino acid combinations which were not experimentally measured, whereas a colored filling indicates variants that have been measured for activity.

Following this first example, a second study on ML-aided directed evolution for stereoselectivity was published in 2018. Interestingly, it builds upon the same experimental platform as the previous example, namely the enantioselectivity of epoxide hydrolase from ANEH.[33] Starting from only nine experimentally evaluated single-point mutants, the researchers built a model and predicted the enantioselectivity of all combinations of these initial changes (2^9). The algorithm, which was used to predict the new sequences, dubbed innov'SAR, was developed by PEACCEL, a France-based biotechnology start-up focusing on enzyme engineering and drug discovery.[34] Innov'SAR only requires sequence information and experimental protein fitness values for training and subsequent inference. Overall, the applied process can be summarized in four steps: 1) The entire protein sequence is encoded based on each amino acid's physicochemical and biochemical properties; 2) from this numerical protein representation, a protein spectrum is calculated through digital signal processing techniques; 3) the protein signals and their respective fitness values are used to construct a regression model; 4) this regression model finally predicts the properties of all possible variant permutations. Applying these steps to the epoxide hydrolase from ANEH led to predicted sequences which, when evaluated experimentally, revealed enzyme variants with improved enantioselectivity.

3. ML-aided Optimization of Enzyme Activity

Complementing the above-described applications of ML to boost enzyme stereoselectivity, we set out to explore algorithm-aided engineering of regioselectivity and activity. Interested in the late-stage functionalization of complex molecules by direct enzymatic CH activation, we explored the potential of Fe(II)/α-ketoglutarate dependent halogenases for the selective halogenation of soraphen A,[35] a potent anti-fungal agent and a target of pharmaceutical interest.[36] We identified a suitable starting sequence capable of catalyzing the desired halogenation reaction in a previously engineered variant of the halogenase WelO5* from *Hapalopsiphon welwitschii* IC-52-2.[37] Notably, we found that while the wildtype enzyme did not accept the bulky substrate, variants that had been specifically engineered to have a broader substrate spectrum exhibited activity.[37] Based on this initial reference and additional docking studies, we selected three critical residues (V81/A88/I161) for complete randomization, *e.g.,* replacement of each amino acid by all other 19 amino acids. As delineated above, the theoretical size of such a library calculates to 20^3. However, due to the redundancy of the genetic code and sampling reasons, the actual screening effort required to cover all combinations exhaustively increases. Specifically, if the screening aim is to cover at least 95% of all encoded variants in a library, a three-fold oversampling should be targeted,[38] challenging experimentalists.

In our halogenase engineering project, we thus opted to explore ML methods to reduce the experimental screening burden and accelerate the identification of beneficial mutations. Notably, our study considered two main engineering objectives: Firstly, we targeted to increase the overall chlorination activity of the enzyme variants, and secondly, we aimed to control the regioselectivity of the halogenation reaction, which would allow the analysis of several derivatized macrolides in structure-function relationship assays.[35]

As a first step, we experimentally confirmed 504 unique halogenase sequence–function pairs, corresponding to 6.3% of the theoretical library. Based on previous applications of ML in protein engineering,[39–41] we then explored the remaining sequence space *in silico*. Toward this goal, we first represented each variant numerically by concatenating the physicochemical and biochemical properties of the amino acids at each mutation site. Multiple amino acid descriptors exist, such as the very comprehensive AAindex[42] or the T-scale descriptor.[43] In our case, combining the T-scale descriptor and selected additional amino acid characteristics[44] produced the best results. With this representation in hand, we trained a Gaussian process. Gaussian processes have received increased attention in the ML community and have also been applied successfully to protein engineering.[39,40] They are accurate and flexible methods for regression and classification and can give a reliable estimate of their own uncertainty. Following training, our model was then used to make activity and regioselectivity predictions on the library's unexplored sequence space. The best-predicted variants were synthesized and experimentally assayed toward their activity and regioselectivity. Gratifyingly, all seven variants predicted towards increased activity performed well, with four halogenases outperforming the previous best variant (Fig. 3). Similarly, the variants predicted towards selectivity exhibited the desired enzyme trait: While seven out of eight produced halogenases showed high selectivity toward the chlorinated soraphen regioisomer **1b**, variant 'LHG' exhibited not only absolute regio-selectivity but also a doubled activity compared to the previous best **1b** producing variant.[37]

Overall, the algorithm-aided evolution process generated halogenase variants capable of synthesizing three distinct chlorinated species from soraphen A and its derivative soraphen C in quantities sufficient for biological testing. In the phenotypic tests, which were carried out on six key pathogens in crop protection, we found that soraphen A derivative **1b** showed an overall better performance than **1a** whereas a chlorinated soraphen C derivative displayed higher species selectivity than the other investigated compounds.[35]

A further successful computational technique in protein engineering is the analysis of protein sequence activity relationships (ProSAR), which has been successfully applied to construct sev-

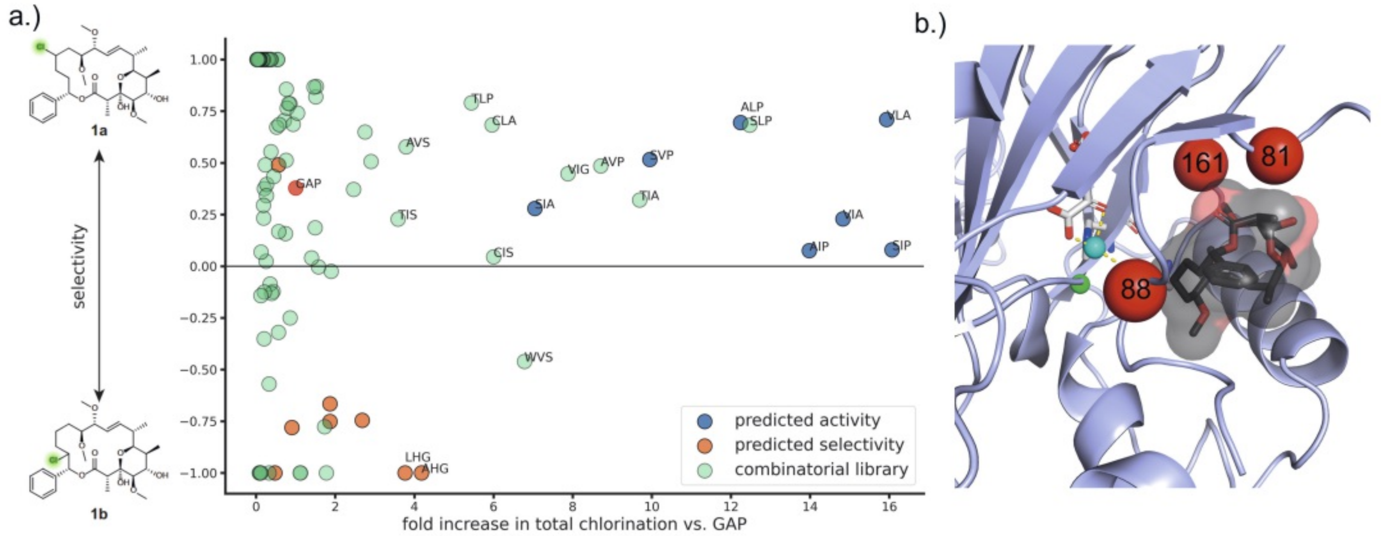


Fig. 3. a) Overview of the experimentally determined activity and regioselectivity results of the three-site combinatorial library of WelO5* (green) and the predicted variants towards activity (blue) and selectivity (orange). Halogenase variants were capable to produce two chlorinated products of soraphen A (**1a** and **1b**). The y-axis shows the regioselectivity of chlorination. The selectivity (S) is calculated using the formula S = $(SIM_{1a} - SIM_{1b})/(SIM_{1a} + SIM_{1b})$. Activity data is normalized to a reference variant (GAP), which was included as an internal reference on each measured 96-well plate. Each variant with a fold-improvement greater than 3.5 is highlighted with a three-letter code representative of the introduced mutations compared to wildtype. For example, V81V/A88L/I161A is shortened to VLA. b) Docking of soraphen A (black) into a model of variant WelO5* V81G/I161P (light blue). The enzyme model was generated using SWISS-MODEL[68] and the crystal structure of WelO5 (PDB ID: 5J4R) as a template. The macrolide soraphen A was docked using AutoDock Vina.[69] The red spheres indicate the targeted positions for the full randomization of the library.

eral highly optimized enzyme variants.[45,46] This technique, which was first published in 2005 by the US-based enzyme engineering company Codexis, facilitated the development of a halohydrin dehalogenase (HHDH) for the industrial production of ethyl (*R*)-4-cyano-3-hydroxybutyrate (HN), improving the enzyme's activity by ~4,000 fold compared to the initial wildtype enzyme (Fig. 4a). To achieve this goal, more than 18 rounds of evolution were carried out, during which 35 distinct mutations were introduced into the wildtype scaffold.[47] In later studies, ProSAR was also employed to increase the stability of a carbonic anhydrase (CA), translating to a 4,000,000-fold improvement over the wildtype in terms of compounded thermostability and alkali tolerance (Fig. 4c).[16] Furthermore, ProSAR enabled the development of a 140,000-fold improved Baeyer-Villiger monooxygenase for the commercial manufacture of esomeprazole used in the blockbuster drug Nexium® by engineering the natural biocatalyst over 19 rounds of evolution.[48] Very recently, ProSAR aided in identifying beneficial mutations in the evolution campaign of an amine transaminase, highly optimized for the efficient production of a chiral sacubitril precursor, a key component of a critical heart failure drug (Fig. 4b).[49]

The multivariate optimization strategy fueling the examples above is an iterative process consisting of diversity generation and statistical modeling. During diversity generation, potentially interesting mutations are generated from various methods, such as rational design, homology modeling, and random mutagenesis. These mutations are then evaluated in combinatorial libraries of varying sizes and screened for activity. A small fraction of this library is sequenced, typically in the order of 3*N, where N is the number of diverse mutations. The generated sequence data then serves as the training set for the statistical analysis. In ProSAR, the statistical modeling step is based on the PLS variable regression technique,[45] which projects the sequence representations to a space of reduced dimensionality to fit a linear model.[50,51] The regression coefficients assigned to each variable represent the impact of a mutation on fitness and are used to decide whether mutations should be retained, discarded, or evaluated again in a different context.[47] Notably, it is not necessarily a priority of ProSAR to find the best variant in each round but rather to rapidly identify beneficial mutations for recombination to reach fitness targets.[16]

As delineated above, the ProSAR-driven approach focuses on parallelized, fast, and efficient iterations in short timeframes. However, not all biocatalysts can be assayed with high throughput at a large scale, and consequently it might be necessary to identify optimal sequences with minimized experimental burden. Such a case was recently described by Greenhalgh *et al.* who targeted an acyl-ACP reductase to produce fatty alcohols *in vivo*.[52] The researchers relied on only 20 sequence–function pairs to initialize an iterative process consisting of *in silico* prediction and experimental evaluations. Rather than predicting which sequences were expected to show the highest activity and evaluating only these variants, the next engineering round was built on an upper-confidence bound criterion. This criterion balances exploration and exploitation,[53,54] by simultaneously exploring areas of uncertainty within the sequence space and assessing possibly improved variants. Such an approach is particularly effective in minimizing the number of evaluations of expensive experiments.[55] The researchers iterated over ten design-test-learn cycles, sampling 10–12 sequences at each iteration, and saw gradual improvements in fatty alcohol titers, cumulating in enzymes that produce above two-fold more fatty alcohols than the wildtype sequences.[52] In our opinion, this Bayesian-type optimization nicely contrasts the ProSAR approach, highlighting how project constraints define the optimization strategy to be used.

4. ML-aided Optimization of Enzyme Stability

Of course, there are other protein properties that researchers attempt to engineer with computational methods, including enzyme stability.[56,57] Notably, a study on the ML-aided engineering of hydrolases for PET depolymerization[58] has recently managed to garner mainstream media attention. Even though more active PET degrading enzymes have previously been developed,[59] the approach is worth highlighting. The involved researchers relied on MutCompute,[60] a 3D self-supervised convolutional neural network, to predict stabilizing mutations. The neural network was trained on a large set of experimentally determined structures from the protein data bank to associate amino acids with neighboring chemical microenvironments with the goal to identify novel gain-of-function mutations.[60] MutCompute was then used to predict which amino acids are not in an optimal configuration for their local environments, effectively performing a single-site saturation scan across all residues in the protein computationally. Sites which the algorithm identified as 'abnormal' were then optimized according to predicted probabilities. This technique was applied to the PET-hydrolysing enzyme (PHE) from *Ideonella sakaiensis* (PETase),[61] and previously engineered variants ThermoPETase[62] and DuraPETase.[63] Validation of the predicted changes revealed scaffolds with improved thermostability (up to 10 °C ΔT_m compared to the respective reference variant), increased protein yield (up to 3.8 fold increase), as well as enhanced catalytic activity (up to 29 fold at selected temperatures).[58]

It should be noted that the MutCompute-type approach is quite different from the examples highlighted above. Rather than learning from a subset of the theoretically available data and predicting fitness within a defined sequence space, biological information is extracted from vast and ever-growing protein databases harnessing the fact that evolution seems to record information about structure and function into evolutionary patterns.[64] This information can be captured, to some extent, by these models and help guide decisions in downstream tasks,[65] complementing and improving the representations used to build models in other machine-learning projects.

5. Conclusion and Outlook

ML is having a notable impact on the biological sciences. Just a few years ago, determining a single protein structure could be a month to year-long process; now, structures can be predicted with similar accuracy within seconds.[64,66] As first engineering examples suggest (*vide supra*), the information contained within the vast sequence and structure datasets already collected might be able to facilitate meaningful predictions even from a few experimentally determined data points. However, not all aspects of protein engineering will benefit equally from ML. The additional costs incurred by sequencing variants, synthesizing the predicted genes, and the time and resources needed to ensure that high-quality data is being provided to train the algorithms must be weighed carefully with the advantages ML provides compared to simply combining beneficial mutations with additive effects.[67] Currently, no clear benchmarks to assess such a benefit exist, as ML accelerated protein engineering examples are scarce, and validating algorithms are restricted to only a handful of datasets.[35,60] Yet, as the field of algorithm-aided enzyme evolution is being more firmly anchored into the biocatalysis sector and gene synthesis and sequencing technologies mature further, we are confident that the *in silico* techniques will evolve into a key element to help address the numbers problem in directed evolution.

Acknowledgements

This work was created as part of NCCR Catalysis, a National Centre of Competence in Research funded by the Swiss National Science Foundation (Grant number 180544).

Received: December 14, 2022

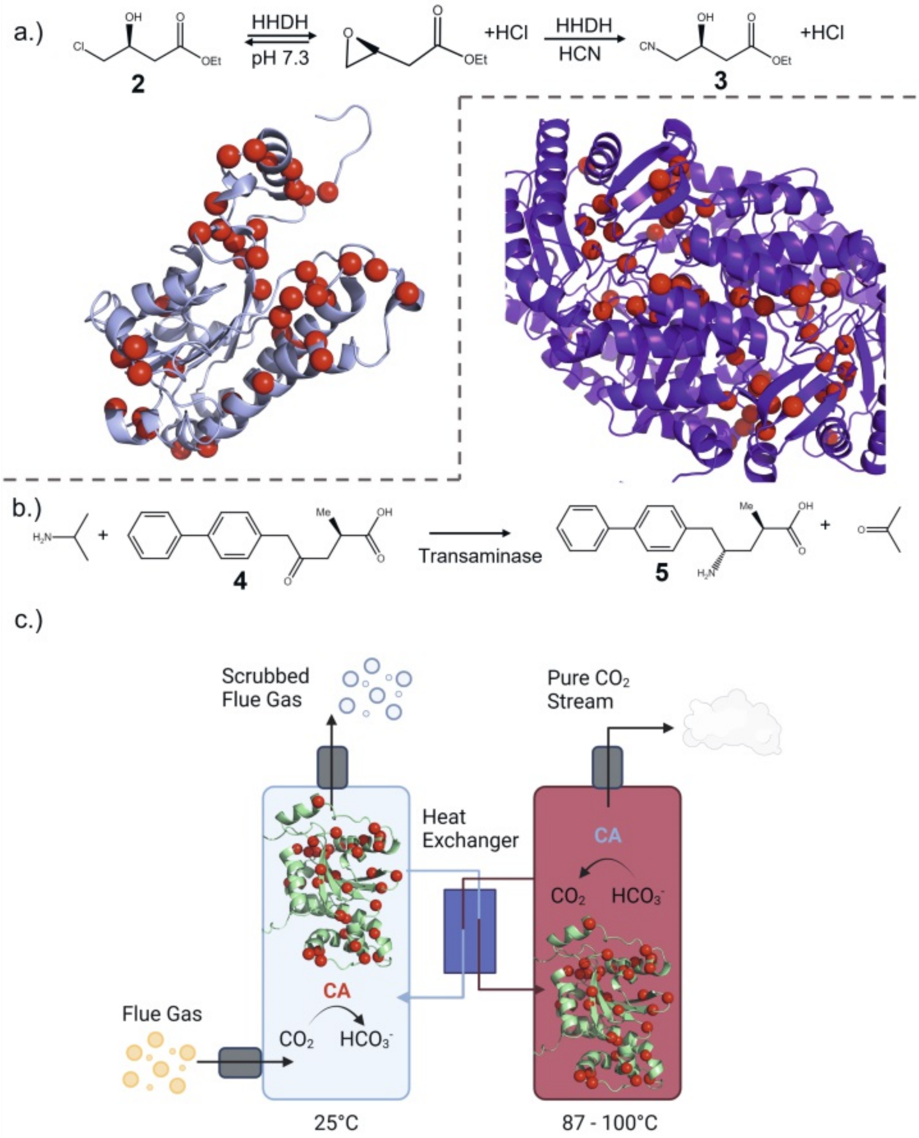


Fig. 4. Overview of successful ProSAR applications. a) HHDH catalyzes a single-vessel enzymatic conversion of ethyl (*S*)-4-chloro-3-hydroxybutyrate (**2**) to ethyl (*R*)-4-cyano-3-hydroxybutyrate (**3**). Variants with ~4,000 fold improvements over wildtype were identified after screening approximately 60,000 variants.[47] The evolved protein structure is depicted as a cartoon and mutated residues are visualized as red spheres. b) Engineering of an amine transaminase for the efficient production of (2*R*,4*S*)-5-biphenyl-4-amino-2-methylpentanoic acid (**5**), a precursor to a critical component in the blockbuster heart failure drug Entresto®. The final transaminase variant, obtained after 11 rounds of evolution, enables an economic conversion of ketone **4** with high yield and purity.[49] The evolved transaminase homodimer is shown as a cartoon with mutated residues highlighted as red spheres. c) An engineered carbonic anhydrase for efficient carbon capture from flue gas. The evolved protein, depicted in green with mutations shown as red spheres, is employed in an absorber column (blue pillar) where CO_2 chemisorbs into an amine solvent. The HCO_3^- containing amine solvent and the evolved enzyme are then transferred to a second column, where CO_2 is stripped at elevated temperatures (red pillar). The depicted carbon capture system represents one of the most challenging industrial environments applied to enzymes.[16] Image created with BioRender.com.

[1] M. Leisola, O. Turunen, *Appl. Microbiol. Biotechnol.* **2007**, *75*, 1225, https://doi.org/10.1007/s00253-007-0964-2.
[2] A. Schmid, J. S. Dordick, B. Hauer, A. Kiener, M. Wubbolts, B. Witholt, *Nature* **2001**, *409*, 258, https://doi.org/10.1038/35051736.
[3] T. J. Magliery, *Curr. Opin. Struct. Biol.* **2015**, *33*, 161, https://doi.org/10.1016/j.sbi.2015.09.002.
[4] S. Wu, R. Snajdrova, J. C. Moore, K. Baldenius, U. T. Bornscheuer, *Angew. Chem. Int. Ed.* **2021**, *60*, 88, https://doi.org/10.1002/anie.202006648.
[5] R. Buller, K. Hecht, M. A. Mirata, H. P. Meyer, in 'RSC Catalysis Series', *Vol. 2018-January*, Royal Society Of Chemistry, **2018**, pp. 3, https://doi.org/10.1039/9781782629993-00001.
[6] K. Hecht, H. P. Meyer, R. Wohlgemuth, R. Buller, *Catalysts* **2020**, *10*, 1, https://doi.org/10.3390/catal10121420.
[7] F. Arnold, *Angew. Chem. Int. Ed.* **2018**, *57*, 4143, https://doi.org/10.1002/ange.201708408.
[8] I. Victorino da Silva Amatto, N. Gonsales da Rosa-Garzon, F. Antônio de Oliveira Simões, F. Santiago, N. Pereira da Silva Leite, J. Raspante Martins, H. Cabral, *Biotechnol. Appl. Biochem.* **2022**, *69*, 389, https://doi.org/10.1002/bab.2117.
[9] S. Lutz, *Curr. Opin. Biotechnol.* **2010**, *21*, 734, https://doi.org/10.1016/j.copbio.2010.08.011.
[10] Y. Wang, P. Xue, M. Cao, T. Yu, S. T. Lane, H. Zhao, *Chem. Rev.* **2021**, *121*, 12384, https://doi.org/10.1021/acs.chemrev.1c00260.
[11] R. Blomberg, H. Kries, D. M. Pinkas, P. R. E. Mittl, M. G. Grütter, H. K. Privett, S. L. Mayo, D. Hilvert, *Nature* **2013**, *503*, 418, https://doi.org/10.1038/nature12623.
[12] F. Meyer, R. Frey, M. Ligibel, E. Sager, K. Schroer, R. Snajdrova, R. Buller, *ACS Catal.* **2021**, *11*, 6261, https://doi.org/10.1021/acscatal.1c00678.
[13] M. Eichenberger, S. Hüppi, D. Patsch, N. Aeberli, R. Berweger, S. Dossenbach, E. Eichhorn, F. Flachsmann, L. Hortencio, F. Voirol, S. Vollenweider, U. T. Bornscheuer, R. Buller, *Angew. Chem. Int. Ed.* **2021**, *60*, 26080, https://doi.org/10.1002/anie.202108037.
[14] M. T. Reetz, L. W. Wang, M. Bocola, *Angew. Chem. Int. Ed.* **2006**, *45*, 1236, https://doi.org/10.1002/anie.200502746.
[15] M. Voss, S. Hüppi, D. Schaub, T. Hayashi, M. Ligibel, E. Sager, K. Schroer, R. Snajdrova, R. M. U. Buller, *ChemCatChem* **2022**, https://doi.org/10.1002/cctc.202201115.
[16] O. Alvizo, L. J. Nguyen, C. K. Savile, J. A. Bresson, S. L. Lakhapatri, E. O. P. Solis, R. J. Fox, J. M. Broering, M. R. Benoit, S. A. Zimmerman, S. J. Novick, J. Liang, J. J. Lalonde, *Proc. Natl. Acad. Sci. USA* **2014**, *111*, 16436, https://doi.org/10.1073/pnas.1411461111.
[17] M. T. Reetz, P. Soni, L. Fernández, Y. Gumulya, J. D. Carballeira, *Chem. Commun.* **2010**, *46*, 8657, https://doi.org/10.1039/c0cc02657c.
[18] N. J. Turner, *Nat. Chem. Biol.* **2009**, *5*, 567, https://doi.org/10.1038/nchembio.203.
[19] J. D. Bloom, S. T. Labthavikul, C. R. Otey, F. H. Arnold, *Proc. Natl. Acad. Sci. USA* **2006**, *103*, 5869, https://doi.org/10.1073/pnas.0510098103.
[20] R. D. Lewis, M. Garcia-Borràs, M. J. Chalkley, A. R. Buller, K. N. Houk, S. B. Jennifer Kan, F. H. Arnold, *Proc. Natl. Acad. Sci. USA* **2018**, *115*, 7308, https://doi.org/10.1073/pnas.1807027115.
[21] M. Reetz, *ChemBioChem* **2022**, https://doi.org/10.1002/cbic.202200049.
[22] B. T. Porebski, A. M. Buckle, *Protein Eng., Des. Select.* **2016**, *29*, 245, https://doi.org/10.1093/protein/gzw015.

[23] Z. Wu, S. B. Jennifer Kan, R. D. Lewis, B. J. Wittmann, F. H. Arnold, *Proc. Natl. Acad. Sci. USA* **2019**, *116*, 8852, https://doi.org/10.1073/pnas.1901979116.

[24] S. P. Singh, A. Kumar, H. Darbari, L. Singh, A. Rastogi, S. Jain, in '2017 International Conference on Computer, Communications and Electronics (Comptelix)', **2017**, pp. 162, https://doi.org/10.1109/COMPTELIX.2017.8003957.

[25] F. Furtado, A. Singh, *Int. J. Res. Ind. Eng.* **2020**, *9*, 84, https://doi.org/10.22105/riej.2020.226178.1128.

[26] B. J. Wittmann, K. E. Johnston, Z. Wu, F. H. Arnold, *Curr. Opin. Struct. Biol.* **2021**, *69*, 11, https://doi.org/10.1016/j.sbi.2021.01.008.

[27] G. Li, Y. Dong, M. T. Reetz, *Adv. Syn. Catal.* **2019**, *361*, 2377, https://doi.org/10.1002/adsc.201900149.

[28] X. Feng, J. Sanchis, M. T. Reetz, H. Rabitz, *Chem. Eur. J.* **2012**, *18*, 5646, https://doi.org/10.1002/chem.201103811.

[29] F. Liang, X. J. Feng, M. Lowry, H. Rabitz, *J. Phys. Chem. B* **2005**, *109*, 5842, https://doi.org/10.1021/jp045926y.

[30] N. Shenvi, J. M. Geremia, H. Rabitz, *J. Phys. Chem. A* **2003**, *107*, 2066, https://doi.org/10.1021/jp021932n.

[31] K. W. Moore, A. Pechen, X. J. Feng, J. Dominy, V. J. Beltrani, H. Rabitz, *Phys. Chem. Chem. Phys.* **2011**, *13*, 10048, https://doi.org/10.1039/c1cp20353c.

[32] K. W. Moore, A. Pechen, X. J. Feng, J. Dominy, V. Beltrani, H. Rabitz, *Chem. Sci.* **2011**, *2*, 417, https://doi.org/10.1039/c0sc00425a.

[33] F. Cadet, N. Fontaine, G. Li, J. Sanchis, M. Ng Fuk Chong, R. Pandjaitan, I. Vetrivel, B. Offmann, M. T. Reetz, *Sci. Rep.* **2018**, *8*, https://doi.org/10.1038/s41598-018-35033-y.

[34] B. Offmann, F. Cadet, P. Charton, WO Patent Appl. No. WO2016166253A1, **2016**.

[35] J. Büchler, S. H. Malca, D. Patsch, M. Voss, N. J. Turner, U. T. Bornscheuer, O. Allemann, C. le Chapelain, A. Lumbroso, O. Loiseleur, R. Buller, *Nat. Commun.* **2022**, *13*, https://doi.org/10.1038/s41467-022-27999-1.

[36] A. Naini, F. Sasse, M. Brönstrup, *Nat. Prod. Rep.* **2019**, *36*, 1394, https://doi.org/10.1039/c9np00008a.

[37] T. Hayashi, M. Ligibel, E. Sager, M. Voss, J. Hunziker, K. Schroer, R. Snajdrova, R. Buller, *Adv. Mater.* **2019**, *131*, 18706, https://doi.org/10.1002/ANGE.201907245.

[38] M. T. Reetz, D. Kahakeaw, R. Lohmer, *ChemBioChem* **2008**, *9*, 1797, https://doi.org/10.1002/cbic.200800298.

[39] P. A. Romero, A. Krause, F. H. Arnold, *Proc. Natl. Acad. Sci. USA* **2013**, *110*, https://doi.org/10.1073/pnas.1215251110.

[40] Y. Saito, M. Oikawa, H. Nakazawa, T. Niide, T. Kameda, K. Tsuda, M. Umetsu, *ACS Synth. Biol.* **2018**, *7*, 2014, https://doi.org/10.1021/acssynbio.8b00155.

[41] T. Vornholt, F. Christoffel, M. M. Pellizzoni, S. Panke, T. R. Ward, M. Jeschek, *Sci. Adv.* **2021**, *7*, 1, https://doi.org/10.1126/sciadv.abe4208.

[42] S. Kawashima, P. Pokarowski, M. Pokarowska, A. Kolinski, T. Katayama, M. Kanehisa, *Nucleic Acids Res.* **2008**, *36*, https://doi.org/10.1093/nar/gkm998.

[43] F. Tian, P. Zhou, Z. Li, *J. Mol. Struct.* **2007**, *830*, 106, https://doi.org/10.1016/j.molstruc.2006.07.004.

[44] Z. O. Ibraheem, R. Abd Majid, S. M. Noor, H. M. Sedik, R. Basir, *Malar. Res. Treat.* **2014**, *2014*, https://doi.org/10.1155/2014/950424.

[45] R. Fox, A. Roy, S. Govindarajan, J. Minshull, C. Gustafsson, J. T. Jones, R. Emig, *Protein Eng.* **2003**, *16*, 589, https://doi.org/10.1093/protein/gzg077.

[46] R. Fox, *J. Theor. Biol.* **2005**, *234*, 187, https://doi.org/10.1016/j.jtbi.2004.11.031.

[47] R. J. Fox, S. C. Davis, E. C. Mundorff, L. M. Newman, V. Gavrilovic, S. K. Ma, L. M. Chung, C. Ching, S. Tam, S. Muley, J. Grate, J. Gruber, J. C. Whitman, R. A. Sheldon, G. W. Huisman, *Nat. Biotechnol.* **2007**, *25*, 338, https://doi.org/10.1038/nbt1286.

[48] Y. K. Bong, S. Song, J. Nazor, M. Vogel, M. Widegren, D. Smith, S. J. Collier, R. Wilson, S. M. Palanivel, K. Narayanaswamy, B. Mijts, M. D. Clay, R. Fong, J. Colbeck, A. Appaswami, S. Muley, J. Zhu, X. Zhang, J. Liang, D. Entwistle, *J. Org. Chem.* **2018**, *83*, 7453, https://doi.org/10.1021/acs.joc.8b00468.

[49] S. J. Novick, N. Dellas, R. Garcia, C. Ching, A. Bautista, D. Homan, O. Alvizo, D. Entwistle, F. Kleinbeck, T. Schlama, T. Ruch, *ACS Catal.* **2021**, *11*, 3762, https://doi.org/10.1021/acscatal.0c05450.

[50] K. K. Yang, Z. Wu, F. H. Arnold, *Nat. Methods* **2019**, *16*, 687, https://doi.org/10.1038/s41592-019-0496-6.

[51] P. Geladi, B. R. Kowalski, *Analytica Chim. Acta* **1986**, *185*, 1, https://doi.org/https://doi.org/10.1016/0003-2670(86)80028-9.

[52] J. C. Greenhalgh, S. A. Fahlberg, B. F. Pfleger, P. A. Romero, *Nat. Commun.* **2021**, *12*, https://doi.org/10.1038/s41467-021-25831-w.

[53] P. Auer, *J. Mach. Learn. Res.* **2002**, *3*, 397, https://doi.org/10.1162/153244303321897663.

[54] N. Srinivas, A. Krause, S. M. Kakade, M. Seeger, in *IEEE Transactions on Information Theory*, vol. 58, no. 5, pp. 3250-3265, May **2012**, https://doi.org/10.1109/TIT.2011.2182033.

[55] J. Snoek, H. Larochelle, R. P. Adams, *Adv. Neural Inf. Process Syst.* **2012**, *4*, 2951, https://doi.org/10.48550/ARXIV.1206.2944.

[56] Y. Li, D. A. Drummond, A. M. Sawayama, C. D. Snow, J. D. Bloom, F. H. Arnold, *Nat. Biotechnol.* **2007**, *25*, 1051, https://doi.org/10.1038/nbt1333.

[57] J. R. Klesmith, J.-P. Bacik, E. E. Wrenbeck, R. Michalczyk, T. A. Whitehead, *Proc. Natl. Acad. Sci. USA* **2017**, *114*, 2265, https://doi.org/10.1073/pnas.1614437114.

[58] H. Lu, D. J. Diaz, N. J. Czarnecki, C. Zhu, W. Kim, R. Shroff, D. J. Acosta, B. R. Alexander, H. O. Cole, Y. Zhang, N. A. Lynd, A. D. Ellington, H. S. Alper, *Nature* **2022**, *604*, 662, https://doi.org/10.1038/s41586-022-04599-z.

[59] V. Tournier, C. M. Topham, A. Gilles, B. David, C. Folgoas, E. Moya-Leclair, E. Kamionka, M. L. Desrousseaux, H. Texier, S. Gavalda, M. Cot, E. Guémard, M. Dalibey, J. Nomme, G. Cioci, S. Barbe, M. Chateau, I. André, S. Duquesne, A. Marty, *Nature* **2020**, *580*, 216, https://doi.org/10.1038/s41586-020-2149-4.

[60] R. Shroff, A. W. Cole, D. J. Diaz, B. R. Morrow, I. Donnell, A. Annapareddy, J. Gollihar, A. D. Ellington, R. Thyer, *ACS Synth. Biol.* **2020**, *9*, 2927, https://doi.org/10.1021/acssynbio.0c00345.

[61] S. Yoshida, K. Hiraga, T. Takehana, I. Taniguchi, H. Yamaji, Y. Maeda, K. Toyohara, K. Miyamoto, Y. Kimura, K. Oda, *Science* **2016**, *351*, 1196, https://doi.org/10.1126/science.aad6359.

[62] H. F. Son, I. J. Cho, S. Joo, H. Seo, H. Y. Sagong, S. Y. Choi, S. Y. Lee, K. J. Kim, *ACS Catal.* **2019**, *9*, 3519, https://doi.org/10.1021/acscatal.9b00568.

[63] Y. Cui, Y. Chen, X. Liu, S. Dong, Y. Tian, Y. Qiao, R. Mitra, J. Han, C. Li, X. Han, W. Liu, Q. Chen, W. Wei, X. Wang, W. Du, S. Tang, H. Xiang, H. Liu, Y. Liang, K. N. Houk, B. Wu, *ACS Catal.* **2021**, *11*, 1340, https://doi.org/10.1021/acscatal.0c05126.

[64] Z. Lin, H. Akin, R. Rao, B. Hie, Z. Zhu, W. Lu, N. Smetanin, R. Verkuil, O. Kabeli, Y. Shmueli, A. dos Santos Costa, M. Fazel-Zarandi, T. Sercu, S. Candido, Alexander Rives, *BioRxiv* **2021**, https://doi.org/10.1101/2022.07.20.500902.

[65] A. Elnaggar, M. Heinzinger, C. Dallago, G. Rehawi, W. Yu, L. Jones, T. Gibbs, T. Feher, C. Angerer, M. Steinegger, D. Bhowmik, B. Rost, *IEEE Trans. Pattern Anal. Mach. Intell.* **2021**, *14*, 1, https://doi.org/10.1109/TPAMI.2021.3095381.

[66] J. Jumper, R. Evans, A. Pritzel, T. Green, M. Figurnov, O. Ronneberger, K. Tunyasuvunakool, R. Bates, A. Žídek, A. Potapenko, A. Bridgland, C. Meyer, S. A. A. Kohl, A. J. Ballard, A. Cowie, B. Romera-Paredes, S. Nikolov, R. Jain, J. Adler, T. Back, S. Petersen, D. Reiman, E. Clancy, M. Zielinski, M. Steinegger, M. Pacholska, T. Berghammer, S. Bodenstein, D. Silver, O. Vinyals, A. W. Senior, K. Kavukcuoglu, P. Kohli, D. Hassabis, *Nature* **2021**, *596*, 583, https://doi.org/10.1038/s41586-021-03819-2.

[67] E. J. Ma, E. Siirola, C. Moore, A. Kummer, M. Stoeckli, M. Faller, C. Bouquet, F. Eggimann, M. Ligibel, D. Huynh, G. Cutler, L. Siegrist, R. A. Lewis, A. C. Acker, E. Freund, E. Koch, M. Vogel, H. Schlingensiepen, E. J. Oakeley, R. Snajdrova, *ACS Catal.* **2021**, *11*, 12433, https://doi.org/10.1021/acscatal.1c02786.

[68] A. Waterhouse, M. Bertoni, S. Bienert, G. Studer, G. Tauriello, R. Gumienny, F. T. Heer, T. A. P. de Beer, C. Rempfer, L. Bordoli, R. Lepore, T. Schwede, *Nucleic Acids Res.* **2018**, *46*, W296, https://doi.org/10.1093/nar/gky427.

[69] O. Trott, A. J. Olson, *J. Comput. Chem.* **2009**, *31*, 455, https://doi.org/10.1002/jcc.21334.

License and Terms

This is an Open Access article under the terms of the Creative Commons Attribution License CC BY 4.0. The material may not be used for commercial purposes.

The license is subject to the CHIMIA terms and conditions: (https://chimia.ch/chimia/about).

The definitive version of this article is the electronic one that can be found at https://doi.org/10.2533/chimia.2023.116

Article II

nature COMMUNICATIONS

ARTICLE

https://doi.org/10.1038/s41467-022-27999-1 OPEN

Check for updates

Algorithm-aided engineering of aliphatic halogenase WelO5* for the asymmetric late-stage functionalization of soraphens

Johannes Büchler[1,2], Sumire Honda Malca[1], David Patsch[1,3], Moritz Voss[1], Nicholas J. Turner[2], Uwe T. Bornscheuer[3], Oliver Allemann[4,5], Camille Le Chapelain[4], Alexandre Lumbroso[4], Olivier Loiseleur[4]✉ & Rebecca Buller[1]✉

Late-stage functionalization of natural products offers an elegant route to create novel entities in a relevant biological target space. In this context, enzymes capable of halogenating sp^3 carbons with high stereo- and regiocontrol under benign conditions have attracted particular attention. Enabled by a combination of smart library design and machine learning, we engineer the iron/α-ketoglutarate dependent halogenase WelO5* for the late-stage functionalization of the complex and chemically difficult to derivatize macrolides soraphen A and C, potent anti-fungal agents. While the wild type enzyme WelO5* does not accept the macrolide substrates, our engineering strategy leads to active halogenase variants and improves upon their apparent k_{cat} and total turnover number by more than 90-fold and 300-fold, respectively. Notably, our machine-learning guided engineering approach is capable of predicting more active variants and allows us to switch the regio-selectivity of the halogenases facilitating the targeted analysis of the derivatized macrolides' structure-function activity in biological assays.

---

[1] Competence Center for Biocatalysis, Institute of Chemistry and Biotechnology, Zurich University of Applied Sciences, Einsiedlerstrasse 31, 8820 Wädenswil, Switzerland. [2] School of Chemistry, The University of Manchester, Manchester Institute of Biotechnology, Manchester M1 7DN, United Kingdom. [3] Institute of Biochemistry, Dept. of Biotechnology & Enzyme Catalysis, Greifswald University, Felix-Hausdorff-Strasse 4, 17487 Greifswald, Germany. [4] Syngenta Crop Protection AG, Schaffhauserstrasse 101, 4332 Stein, Switzerland. [5] Present address: Idorsia Pharmaceuticals Ltd, Hegenheimermattweg 91, 4123 Allschwil, Switzerland. ✉email: olivier.loiseleur@syngenta.com; rebecca.buller@zhaw.ch

Subtle molecular changes in small molecules can have a profound impact on their biological activity and metabolism. For example, monodechlorinated and didechlorinated vancomycin lose approximately 30 and 50% of the antimicrobial effect exhibited by parent antibiotic vancomycin[1], respectively. Similarly, the introduction of a single methyl group led to MK-8133, a dual orexin receptor antagonist, with 480-fold boosted potency[2]. In the latter example, the methyl group had to be installed through a laborious five-step *de novo* synthesis[3]. In contrast, late-stage functionalization (LSF) of C–H bonds offers direct access to new analogs of a lead structure. In this way, LSF constitutes a valuable tool to investigate structure-activity relationships of small molecules, especially natural products[4], and supports the optimization of on-target potency, selectivity, and absorption-distribution-metabolism-excretion (ADME) properties while helping to improve physical properties such as solubility and stability. In addition, LSF can be of aid in the protection and exploration of novel intellectual property space by giving access to molecular entities left unexplored by conventional synthetic approaches[3]. Typical functionalizations of C–H bonds include oxygenation, amination, methylation, borylation, thianthrenation, azidation, and halogenation[3,5]. Notably, the incorporation of chlorine and bromine offers new routes to modify the molecule through cross-coupling chemistry or substitution reactions[6].

The synthesis route to organohalides commonly involves multiple steps. In order to achieve high chemo-, regio- and stereoselectivities[7], the use of protecting, directing, or activating groups is often necessary. As some of these groups may need to be removed in subsequent steps, such approaches lack atom economy. Overall, the halogenation of unactivated C–H bonds remains a challenge for chemists[8,9]. Enzymatic halogenations, on the other hand, often exhibit excellent regio- and stereoselectivity even in complex molecular settings, therefore complementing—and sometimes outperforming— existing strategies[10–13].

Biocatalytic halogenations are carried out by enzymes called halogenases, which are typically classified according to their catalytic mechanism: Heme, vanadium, and flavin-dependent halogenases (Fl-Hals) follow an electrophilic aromatic substitution mechanism via the generation of hypohalous acid, iron/α-ketoglutarate dependent halogenases (αKGHs) employ a radical pathway, while *S*-adenosyl-L-methionine (SAM) fluorinases react via a nucleophilic substitution[14]. In contrast to the electrophilic Fl-Hals, which act on electron-rich *sp*^{2}-carbons through the intermittent generation of hypohalous acid, αKGHs can functionalize unactivated C(*sp*^{3})-H bonds. The catalytic mechanism is based on the generation of a high-valent $Fe^{IV}=O$ intermediate capable of abstracting a hydrogen atom from the substrate. The resulting carbon radical is then coupled to the iron-coordinated chlorine, thereby affording the corresponding halogenated compound in a regio- and stereoselective manner (Fig. 1a). In recent years, a handful of αKGHs have been described: The carrier-protein dependent halogenases BarB1 and BarB2[15], SyrB2[16], CytC3[17], CmaB[18], HctB[19], CurA[20] and the synthetically more interesting freestanding halogenases WelO5[21], WelO5*[22], *Wi*-WelO15[23], AmbO5[24], the BesD[25] family, the recently identified plant halogenases SaDAH and McDAH[26] as well as the halogenase AdeV[27], which acts on nucleotide substrates.

To date, halogenase engineering has mainly focused on Fl-Hals[10,28–33] or haloperoxidases[34,35] with the aim to provide catalysts capable to derivatize non-natural substrates en route to more valuable aryl-, alkoxy or amino acid compounds[36–40] or for their use as final products[41,42]. In contrast to the wealth of reports on Fl-Hals, the number of αKG-dependent halogenases is small and their reported substrate scope is mainly limited to their natural substrates and close analogs. In 2019, the first examples of engineering freestanding αKGHs toward non-natural substrates were reported by us and others[23,43]. The studies highlighted the malleability of αKGHs WelO5* and *Wi*-WelO15 by tailoring the enzymes for the regio- and stereoselective chlorination of a non-alkaloid type substrate and more closely related substrate analogs of 12-*epi*-hapalindole C, respectively. In both cases, substantial increases in apparent k_{cat} (WelO5*: 400-fold compared to wild type; *Wi*-WelO15: 276-fold compared to first-generation mutant) could be achieved by enzyme engineering[23,43]. Despite the pioneering nature of these engineering studies, it should be noted, that the chosen non-natural substrates were similar in size and shape to the halogenases' natural substrate 12-*epi*-hapalindole C.

Soraphens are the largest known family of myxobacterial polyketides and display a diverse array of chemical moieties (e.g., unsubstituted phenyls and sensitive allylic ethers amongst other features) which render them an attractive test case for an application to a broader range of polyketides. Soraphen A, the main representative of the soraphens, was identified in the supernatant of the *Sorangium cellulosum* strain Soce26 and shows inhibitory activity against several phytopathogenic fungi through inhibition of acetyl-coenzyme A carboxylase (ACC)[44]. The crystal structure of the yeast biotin carboxylase (BC) domain complexed with soraphen A (PDB ID: 1W96) revealed that the macrolide acts as an allosteric inhibitor[45] by disrupting dimerization of the BC domain and stabilizing the catalytically inactive monomer (Supplementary Fig. 1)[46]. Even though highly potent, the further development of these natural products as potent antifungal agents has been hampered due to off-target selectivity concerns and sensitization in mammals[47]. Notably, soraphen A has recently also become a target of pharmaceutical interest[44]. In cancer therapy research, several studies established that tumoral cells have a dependence on *de novo* fatty acid synthesis and that inhibition of ACC triggers apoptosis with no or little effects on healthy cells[48]. Modified lead structures are therefore sought after, both in agrochemistry as well as in pharmaceutical chemistry, which—owing to the complexity and sensitivity of the natural product—are, however, difficult to obtain in the quantities and within the timeframes required by modern drug discovery[49]. Consequently, the development of adapted synthetic methodologies, including biocatalytic transformations, are of key interest to drive the development of complex, natural compounds into useful products.

In this work, we assess the biocatalytic potential of αKGHs by employing algorithm-assisted enzyme engineering to tailor the recently described non-heme iron halogenase WelO5* from *Hapalosiphon welwitschii* IC-52-3 for selective halogenation of soraphen A (**1**), soraphen C (**2**) and their semi-synthetic analogs **3** and **4** (Fig. 1b). Phenotypic testing of the derivatized macrolides against six different fungal key pathogens in crop protection is carried out to inform about the halogenated macrolides' biological activity.

Results

Synthesis of starting material. Soraphen structures contain ten stereocenters, including hydroxyl-, methyl, methoxy, and a hemiacetal group rendering these natural products biologically highly interesting but chemically very complex molecules. In addition, such polyketide macrocycles are also known to adopt several conformations[50]. While soraphen A can be accessed through an optimized bioprocess[47,51], its penultimate biosynthetic congener soraphen C is a much less explored member of the soraphen family and very difficult to isolate in sufficient amounts from fermentation despite its value as a chemical probe[52]. To obtain the compound for our study, we, therefore, developed a concise semisynthesis starting from soraphen A

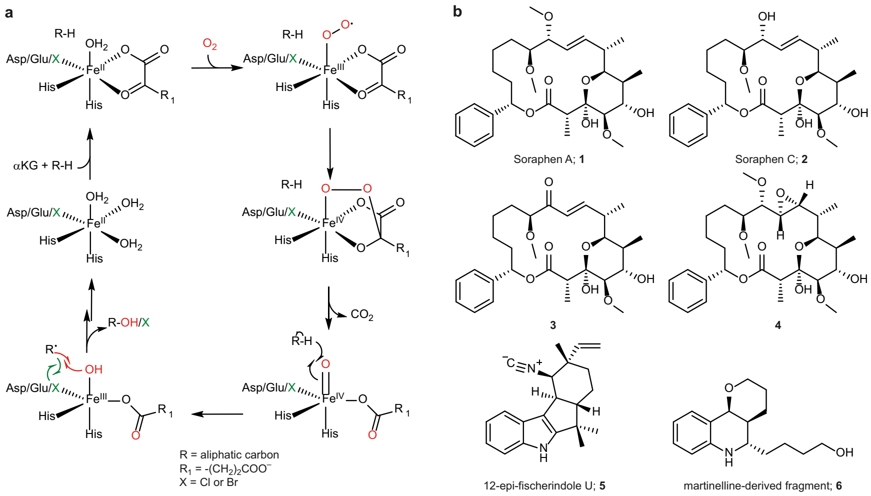


Fig. 1 Proposed reaction mechanism and substrates of wild type and engineered WelO5* variants. a Proposed reaction mechanism of Fe(II)/αKG-dependent halogenases and hydroxylases. Mechanism adapted from Mitchell et al.[66] and Galonic et al.[79] **b** Structural comparison of the macrolide soraphen A and its analogs (**1-4**) with WelO5*'s natural substrate 12-epi-fischerindole U (**5**)[22] and the accepted martinelline-derived fragment (**6**)[43].

(Supplementary Fig. 2). This route, entailing selective oxidative demethylation of the allylic methoxy group and a subsequent stereo-directed reduction of the intermediate ketone, offers the first synthetic access to soraphen C. Even though soraphen C had been obtained earlier through fermentation[53], we are now reporting the first complete characterization of this natural product.

Identification of an active starting halogenase for halogenation of soraphen A. To identify a halogenase which would accept soraphen A, an enzyme panel consisting of 59 native and engineered electrophilic and freestanding aliphatic halogenases capable of acting on a wide range of sp^2 and sp^3-carbons was screened (Supplementary Tables 1, 2). The engineered Fl-Hals included in the panel were derived from literature[31,41,42,54] whereas the engineered αKGHs consist of WelO5* variants which we had previously identified as possessing a broadened substrate scope[43].

All halogenases, expressed in *E. coli* BL21(DE3), were used for crude cell-lysate biotransformations of soraphen A in a deep-well plate. While neither halogenation nor hydroxylation activity toward the target substrate was detected for any of the wild type enzymes, liquid-chromatography coupled to mass spectrometry (LC-MS) analysis showed that biotransformations with 26 out of the 28 included WelO5* variants led to the formation of derivatized soraphen A. In particular, variants V81G/I161P, V81G/I161G, as well as I161A, showed appreciable activity leading to the detection of three prominent products with *m/z* ratios of 577.2 and 559.2, which are consistent with the calculated mass of two chlorinated products and a hydroxylated product, respectively (Supplementary Fig. 3). The structures of the chlorinated products **1a** and **1b**, as well as the hydroxylated product **1c**, were solved using nuclear magnetic resonance (NMR) analysis, which confirmed chlorination and hydroxylation of aliphatic carbon centers of **1** (Supplementary Fig. 4). Notably, the enzymatic derivatization occurred at positions in the molecule which would have been difficult to target via traditional chemical means and opens options for further functionalization in previously unexplored segments of the molecule.

In contrast to previous engineering studies on WelO5*, the reaction selectivity (halogenation vs. hydroxylation) of the best-performing variant V81G/I161P was slightly in favor of the halogenation reaction (2:1 halogenation to hydroxylation ratio, estimated via the SIM areas of the product peaks). This is remarkable for the transformation of a structurally highly divergent compound compared to the natural substrate 12-*epi*-fischerindole U (Fig. 1b). As has been observed for WelO5* and other αKGHs, this enzyme family's reaction selectivity is strongly governed by the substrate structure and substrate positioning in the active site. It has been shown that the biotransformation of 12-*epi*-hapalindole C, another literature-known native substrate of WelO5* similar in structure to 12-*epi*-fischerindole U[22], led to the predominant formation (ca. 50%) of hydroxylated product and 25% of the desired chlorinated product 12-*epi*-hapalindole E[43]. Other examples include studies on the carrier-protein-dependent halogenase SyrB2, which turned into an effective hydroxylase in response to the length of added C-atoms in its native substrate L-threonine[55].

Enzyme engineering of WelO5* for improved activity and selectivity. While wild type WelO5* did not accept soraphen A, mutation of only two residues near the active site conferred initial halogenation and hydroxylation activity toward the bulky macrolide substrate. This activity data underlines the striking malleability of WelO5*[23,43] allowing a considerable expansion of substrate scope by exchange of very few amino acids strategically positioned in the vicinity of the reactive iron species. Docking of soraphen A into a model of the best-performing WelO5* variant V81G/I161P, which was created using SWISS-MODEL[56], led to solutions in which the active site was capable to accommodate soraphen A (Fig. 2). Based on these docking results and in agreement with the studies from Hayashi et al.[43] and Duewel et al.[23], three critical amino acid positions, namely 81, 88, and 161 (Fig. 2), were chosen for full randomization in a library targeted for the use in an algorithm-aided enzyme engineering strategy.

Traditionally, gene mutagenesis methods for the generation of variant libraries are PCR-based techniques and include error-prone polymerase chain reaction (epPCR), saturation mutagenesis, or DNA shuffling. Saturation mutagenesis, as required in our approach, is a highly advantageous technique in rational enzyme design, however, it is known to suffer from amino acid bias leading to reduced library quality and thus increased screening effort[57]. In order to allow for an unbiased library construction, we opted for a *de novo* library synthesis using high-fidelity on-chip solid-phase gene synthesis[58]. This library construction strategy allowed us to limit library diversity to the theoretical 8000

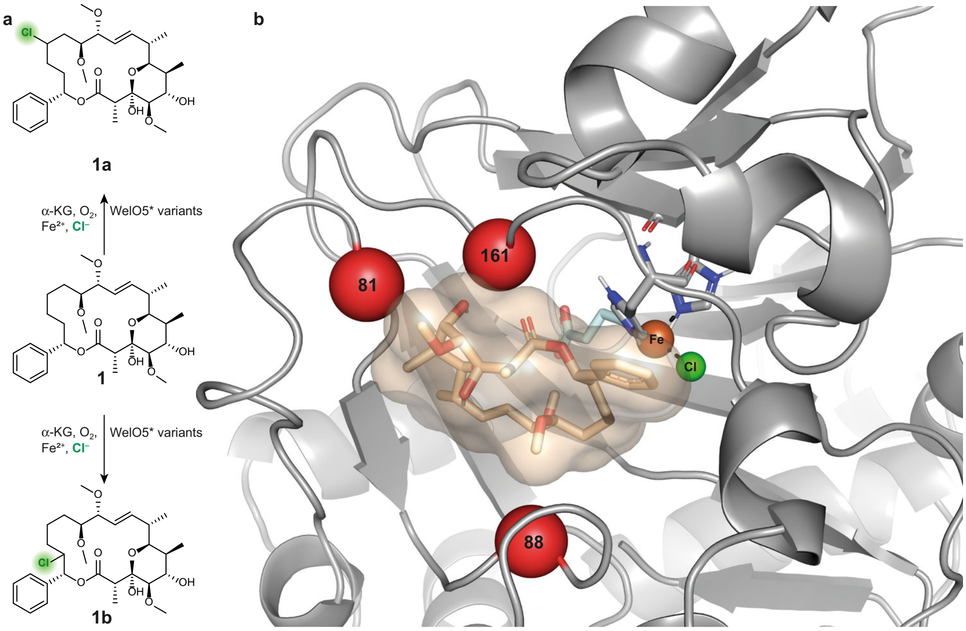


Fig. 2 Identification of target sites for enzyme engineering. **a** Regio-divergent halogenation of soraphen A in function of the employed WelO5* variant. **b** Docking of soraphen A (wheat) into a model of variant WelO5* V81G/I161P (gray). The enzyme model was created using SWISS-MODEL[56] and the crystal structure of WelO5 (PDB ID: 5J4R) as a template. Soraphen A was docked using AutoDock Vina[77]. The red spheres indicate the targeted positions for the full randomization library. Histidine residues and the α-ketoglutarate (pale cyan) in complex with the iron (orange sphere) are shown as sticks. The chlorine coordinating to the iron is shown as a green sphere.

variants (20^3) for the full co-randomization of residues at positions 81, 88, and 161 and minimize screening effort. For simplicity, we will report WelO5* variants with a three-letter code hereafter. For instance, wild type WelO5*, which contains the amino acids V81/A88/I161, is denoted as variant VAI, whereas variant V81G/A88A/I161P, which was identified as being active on soraphen A in the initial hit panel screening, is dubbed GAP.

The synthetic gene library was ordered from Twist Bioscience. The gene fragments were cloned into the pET28b(+) vector and transformed into *E. coli* BL21(DE3) cells in house. About 504 unique variants (6.3% of the theoretical library) were confirmed by Sanger sequencing and screened for the derivatization of soraphen A (Fig. 3, red circles). As expected, we observed the formation of the previously identified products in addition to a second hydroxylated compound (**1d**). Overall, four distinct soraphen A analogs could be produced by the analyzed enzyme variants: Chlorination products **1a** and **1b**, as well as hydroxylation products **1c** and **1d**, were observed (Supplementary Fig. 4). In all cases, hydroxylation product **1d** was a side product and formed only in minimal amounts (max. formation of 2%, not isolated).

In comparison to the previously best-performing variant GAP, we identified amino acid combinations (VIG, AVP, and TIA) that boosted total chlorination activity for soraphen A by 8-10-fold, whereas variant SLP increased the total halogenation activity by 13-fold. In addition to improving total chlorination activity, the three-site combinatorial library also contained variants, which modulated the regioselectivity of the halogenation reaction. Instead of preferentially forming product **1a**, variant LHS exclusively led to chlorination product **1b** while remaining similar in total chlorination activity to variant GAP.

While the theoretical number of unique variants in a *de novo* synthesized three-site combinatorial library is 8000, a much higher number of samples will have to be screened in practice. This is because the degree of oversampling increases with the percentage of targeted library coverage. As a result, a library coverage of 95% will require the analysis of ~24,000 variants[59], an effort which demands considerable resources. Inspired by previous successful applications of machine learning in protein engineering[60–62], we explored the remaining protein landscape in silico using Gaussian processes, allowing us to reduce the physical screening burden and accelerate the accumulation of beneficial mutations. By representing amino acids as a 17-dimensional vector, which was obtained by concatenating the five-dimensional T-scale descriptor[63] and additional amino acid characteristics[64], our machine learning approach then defined the feature vector of a sequence by joining the vector representation of its individual amino acids at sites V81X, A88X, and I161X. With this strategy, we were able to identify both more active and more selective variants with noticeable accuracy and precision (Supplementary Fig. 5 and Supplementary Table 3). All seven variants predicted towards activity (Fig. 3, blue circles) were highly active, with four of them outperforming the previous best variant SLP (up to a 16-fold increase over GAP). Predictions toward selectivity (Fig. 3, green circles) show a similarly high fraction of improved variants, with one of them enhancing activity over the previously most selective variant for chlorination site B by >2-fold while retaining a complete selectivity for regioisomer **1b**.

While the detailed mechanism behind the improved activity of the evolved variants remains unclear, we attempted to get a better understanding of the factors governing the regioselectivity of the evolved variants by carrying out docking studies with substrate **1**. For these experiments, we used the available crystal structure of WelO5 (PDB ID: 5J4R), a close homolog of WelO5*, as a basis of our homology modeling with the tool SWISS-MODEL[56]. Comparing the docking results of variant GAP, our most selective variant for the production of **1a**, with the analysis of variant AHG, our most selective variant for the synthesis of **1b**, we observed a shift in substrate positioning with respect to the iron-oxo and the Cl-ligand (Fig. 4 and Supplementary Fig. 6). The set of mutations acquired in AHG presumably changes the binding

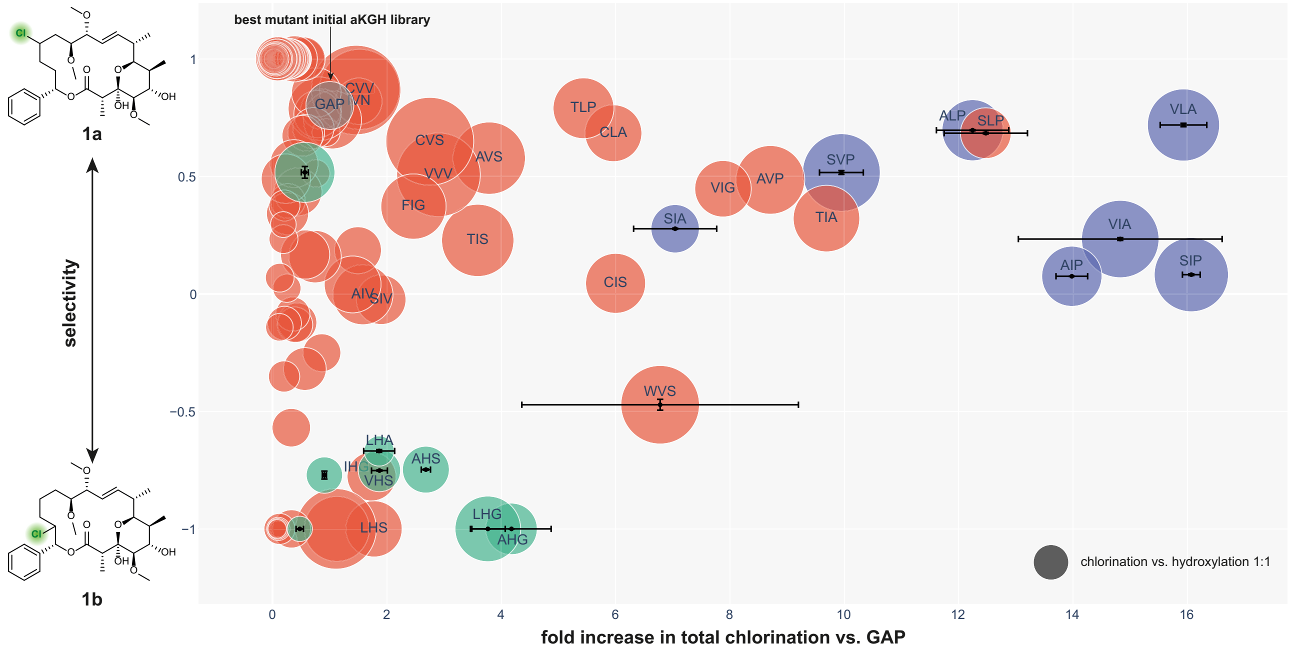


Fig. 3 Biotransformation results of the combinatorial library of WelO5* (red) and the predicted variants (blue and green). Two chlorinated products (**1a**, **1b**) of soraphen A were detected. The y-axis shows the regioselectivity of the chlorination. The selectivity (S) is calculated using the formula $S = (SIM_{1a} - SIM_{1b})/(SIM_{1a} + SIM_{1b})$. For variants showing a higher than 1.5-fold increase in total chlorination compared to WelO5* V81G/A88A/I161P (GAP, gray) the amino acid sequence of the three engineered positions is shown. On each measured 96-well plate, the variant GAP and negative controls were included as internal references. The combinatorial library variants were measured once. Predicted variants (selectivity; activity) and best-performing variants (SLP; WVS) were measured in triplicate as individual experiments. Data were presented as mean values ± standard deviation. The size of the circle corresponds to higher chlorination vs hydroxylation activity. As a reference, the dark gray sphere corresponds to a 1:1 halogenation to hydroxylation activity (area halogenation products/area hydroxylation products).

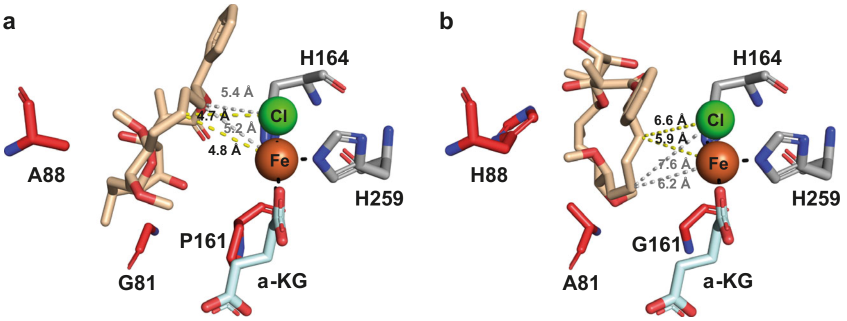


Fig. 4 Soraphen A (wheat) docked models of the regio-divergent WelO5* variants GAP and AHG. a In the model of WelO5* variant GAP (amino acids G, A, and P shown as red sticks) shorter distances between the iron and chloride to C14 of soraphen A (yellow dotted lines) than to C16 of the macrolide (gray dotted lines) suggest the structural reason for the predominant formation of regioisomer **1a**. **b** In the soraphen A docked model of WelO5* variant AHG (amino acids A, H, and G shown as red sticks), a shift in substrate positioning leads to shorter distances between the iron and chloride to C16 of soraphen A (yellow dotted lines) than to C14 of the macrolide (gray dotted lines) underlining the observation of selectivity for formation of regioisomer **1b**. Histidine residues (gray) and the α-ketoglutarate (pale cyan) in complex with the iron (orange sphere) are shown as sticks. The chlorine coordinating to the iron is shown as a green sphere.

mode of soraphen A in such a way, that H-abstraction is now favored from a different C–H bond, namely C16, instead of C14 as observed for GAP (Supplementary Fig. 7).

To further assess the substrate promiscuity of the engineered WelO5* variants and to expand our palette of uniquely derived soraphen analogs for biological testing, we analysed the transformation of soraphen C (**2**) and the soraphen analogs **3** and **4**. In analogy to soraphen A, we observed the formation of two chlorinated and hydroxylated products for soraphen C. Also, the soraphen analogs **3** and **4** led to the formation of several singularly derivatized macrolide structures (Supplementary Table 4). Interestingly, the initial whole-cell screening using soraphen A as a substrate did not reveal doubly chlorinated or doubly hydroxylated products nor a mixture thereof. To further investigate the substrate promiscuity of our engineered variants, we continued by carrying out in vitro studies applying optimized reaction conditions using mono-chlorinated **1a**, **1b**, and **2a** as substrates and purified enzyme preparations of variants GAP, SLP, VLA, and WVS. Of all combinations tested, variants WelO5* SLP and VLA exhibited detectable substrate promiscuity and proved capable to produce minor amounts of doubly chlorinated products starting from **1a** (0.04% conversion with SLP) and **1b** (1.7% with SLP and 1.9% conversion with VLA) as well as a hydroxylated product derived from **1b** (3.7% with SLP

Table 1 Biochemical characterization of selected WelO5* variants for the biocatalytic production of 1a.

Variant	app. k_{cat} (min^{-1})	app. K_m (mM)	app. k_{cat}/K_m (min^{-1} mM^{-1})	rel. k_{cat}	TTN[+]
GAP	0.026 ± 0.007	0.45 ± 0.14	0.07 ± 0.21	1	0.3 ± 0.2
SLP	2.413 ± 0.349	0.42 ± 0.09	5.74 ± 0.07	93	30.0 ± 8.3
VLA	1.959 ± 0.509	0.44 ± 0.03	4.45 ± 0.07	75	91.8 ± 22.0

[+](TTN experiments were performed in two test series (biological replicates) and each series consisted of four independent experiments ($N = 4$); kinetic parameters are given as the average of $N = 3$ ± SD).

and 5.8% conversion with VLA) (Supplementary Fig. 8). Overall, and in alignment with the observations made for the halogenation of a martinelline-derived fragment by Hayashi et al.[43], the main detectable products of the engineered WelO5* variants under standard reaction conditions were the mono-derivatized soraphens.

Biochemical characterization of improved WelO5* variants. Following our enzyme engineering campaign, we explored the biochemical characteristics of our evolved halogenase variants. For variant GAP, our best initial hit, as well as for variants SLP and VLA, the most active variants for the biocatalytic production of **1a**, Michaelis–Menten kinetics were recorded (Table 1). As initial velocities decreased for all variants with increasing substrate load, a substrate inhibition model was used (Supplementary Eq. 1) to determine the kinetic parameters. Substrate inhibition is a common phenomenon in enzymology and is well documented for enzymes following a radical reaction mechanism. P450 enzymes, for example, have been shown to suffer from decreased activity at high substrate concentrations in function of the provided substrate[65]. Similarly, WelO5* kinetics seems to be governed by the substrate type: While non-classical Michaelis–Menten kinetics were observed when the engineered WelO5* variants were presented with the macrolide soraphen A, the martinelline-derived fragment **6** used in a previous study[43] did not elicit observable substrate inhibition in closely related WelO5* enzyme variants even at concentrations as high as 2.0 mM.

Analysis of the kinetic parameters revealed that variant VLA (apparent $k_{cat} = 1.96\ \text{min}^{-1}$; TTN = 91.8) exhibited a > 75-fold improved apparent k_{cat} and a > 300-fold increased total turnover number (TTN) yielding substantially improved concentrations of product **1b** (Supplementary Fig. 9) when compared to the initial hit, variant GAP (apparent $k_{cat} = 0.03\ \text{min}^{-1}$; TTN = 0.3). Strikingly, engineered VLA displays a similar apparent k_{cat} and total turnover number for the bulky macrolide soraphen A as wild type halogenases acting on their native substrates[12]: wild type WelO5, for example, is reported to halogenate its native substrate 12-*epi*-fischerindole U with a k_{cat} of 1.8 min^{-1} whereas the total turnover number is reported to be 70[24].

It was previously shown that WelO5*[43] and other αKGHs of bacterial[25,66,67] and plant origin[26], can install alternative anions. We, therefore, tested the ability of our best WelO5* variants (GAP, SLP, and WVS) to generate additional soraphen A derivatives using a panel of alternative anions, namely F^-, Br^-, I^-, N_3^-, and NO_2^-. Among the anion tested, Br^-, N_3^-, and NO_2^- were incorporated into the substrate as shown by the appearance of up to two products with the expected *m/z* ratios in selected ion monitoring (Supplementary Fig. 10), in analogy to the product pattern in the corresponding chlorination reactions. Incubation with iodide and fluoride under standard reaction conditions did not yield derivatized product likely due to steric and electronic reasons. As previously observed for WelO5* variants[43] and the freestanding plant halogenase SaDAH[26], the chloride and azide anion yielded the best transformation results as deduced from SIM peak areas. The regioselectivity of alternative anion incorporation was not determined directly. Interestingly, however, distribution between the two observed products when incubating halogenases SLP and WVS with alternative anions reflected the observed product distribution in chlorination reactions leading us to postulate installation of bromide, azide, and nitrate at the same sites in the substrate molecule.

Biological activity of soraphen derivatives against phytopathogenic fungi. Next, we embarked on the biological characterization of the halogenated products. Toward this goal, the biotransformations of soraphen A and soraphen C were carried out at preparative scale (100 mg scale) using the optimized WelO5* variants VLA (soraphen A, halogenation product **1a**), WVS (soraphen A, halogenation product **1b**), and VAA (soraphen C, halogenation product **2a**). In all cases, enough product was obtained and submitted to biological activity profiling. The performed biological tests were phenotypic, i.e., carried out on living fungi, either with a fungal liquid culture or as a preventative application on leaf disk, and considered not only on-target potency but also metabolism, physicochemical properties (leaf penetration for instance), UV stability, and phytotoxicity. The activity is reported as BP80 (break point 80%), which corresponds to the concentration above which 80% of activity, measured as fungal growth inhibition, is observed (Fig. 5, Methods in SI). Six different fungi were evaluated (Fig. 5), as they represent key pathogens in crop protection and cause a large spectrum of crop diseases: *Puccinia recondita* (black rust), *Septoria tritici* (leaf blotch), *Erysiphe graminis* (also called *Blumeria graminis*, powdery mildew), and *Monographella nivalis* (snow mold) attack cereals, especially wheat, while *Botrytis cinerea* (gray mold) acts on horticultural crops including wine grapes, and *Mycosphaerella arachidis* (leaf spots) affects peanut plants. Finding natural molecules to fight these plant pathogens is of special relevance for Europe, where the European Green Deal[68] has become a driver for use of natural products in crop protection.

The aliphatic region of soraphen A, which was derivatized in our experiments, is known to make hydrophobic contact with the acetyl-coenzyme A carboxylase BC domain (in particular with W487, using numbering from PDB ID: 1W96). This critical tryptophan residue is highly conserved within the acetyl-coenzyme A carboxylase BC domain across the tested fungal species. Therefore, the conformational changes in the soraphens, which the chlorine or hydroxyl-group introduction was expected to induce, may also have resulted in a binding penalty which could have led to the observed reduced activity, specifically in the case of hydroxylated compound **1c**. Remarkably, though, all chlorinated analogs conserved a good level of activity on most fungal pathogens, which is unprecedented to date in the ensemble of derivatives accessible from the fully functionalized natural product[47].

As the biological tests performed were phenotypic, a target-based SAR analysis cannot fully explain the activity observed in vivo, which depends on many other factors such as in planta

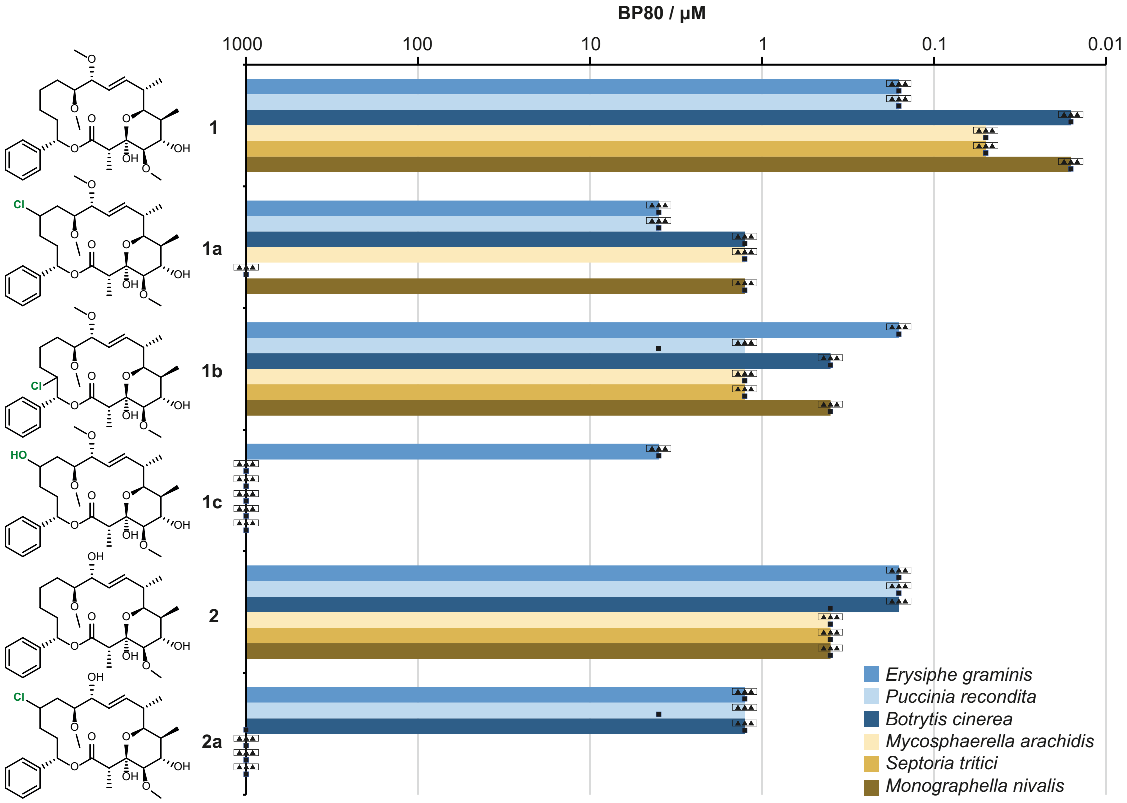


Fig. 5 Break point of efficiency 80% (BP80). Soraphen A (**1**), C (**2**), and enzymatically derivatized analogs (**1a**, **1b**, **1c**, **2a**) were tested against six different fungi, determined by dilution series at four concentrations and measured against positive and negative standards. BP80 represents the concentration of active ingredients above which 80% or more of efficiency is observed. The experiments on living organisms were performed in two test series (biological replicates). The first series consisted of a single experiment ($N = 1$, square) for each fungus, the second series consisted of triplicates with three samples tested in parallel during the same test session ($N = 3$, triangles).

and in fungi metabolism, cell penetration, distinct physicochemical properties of the compounds as well as on differential metabolism and even variations in plant-fungi interactions. Nevertheless, it is worth noting that the site of chlorination seems to impact observed biological activity, **1b** showing an overall better performance than **1a** whereas the chlorinated soraphen C derivative **2a** seems to display higher species selectivity than the other investigated compounds.

Altogether, the observed modulation of the soraphens' biological activity highlights the value of the enzymatic late-stage functionalization approach to generate knowledge in regions of the natural product structure very difficult to access by any chemical means. In fact, spanning over more than 30 years, comprehensive derivatization efforts on the soraphens, which aimed to evaluate whether modified structures might retain good bioactivity, failed: Even minor structural changes led to complete loss of potency[47]. In this context, the activity observed for the here reported chlorinated soraphen analogs and the relatively short time, in which they were obtained especially when compared to total or semisynthesis approaches is even more remarkable. These results represent a good starting point for further structure-activity studies of this class of macrolides and underline the ability of engineered WelO5* halogenases to display unique distance and geometry-based control of functionalization in complex molecules.

Discussion

Here, we demonstrate that through the application of algorithm-assisted enzyme evolution, we endowed WelO5* variants with the capability to halogenate the bulky non-natural substrate soraphen A. Our most active engineered variant WelO5* VLA catalyzes the halogenation of the macrolide **1** to yield product **1a** with an apparent k_{cat} value and a total turnover number which mirror the activity of wild type aliphatic halogenases for their natural substrate (*vide infra*)[12] thus highlighting the malleability of WelO5*'s active site and underlining the effectiveness of our engineering strategy.

Following the identification of hot spots through rational enzyme design, the use of machine learning enabled us to successfully navigate the sequence-function space of a 20^3 combinatorial library of aliphatic halogenase WelO5*. By providing a homogenous and consistent data set of high quality for training and validation of the algorithms, we were able to reliably predict functional properties such as activity and regioselectivity of the enzyme variants from sampling only 6% of the theoretical data points. To date, there are only a few examples that showcase the use of machine learning to improve an enzyme's activity[69,70], and the extent of sampling to obtain predictions varies strongly (Supplementary Table 5). To mature the field, further experimentally confirmed examples such as this one will be necessary to develop more standardized guidelines for the use of machine learning in enzyme engineering and enable comparison between predictors[69]. In addition, the implementation of molecular dynamics simulations into the enzyme engineering workflow might help to further fine-tune machine learning algorithms and —as automation hardware and library design strategies are similarly maturing—allow to interrogate sequence space even more effectively.

Through our resource-saving evolution process, we generated halogenase variants capable of functionalizing soraphen A and soraphen C yielding three distinct halogenated species in quantities sufficient for biological testing. Notably, the enzymatically derivatized positions would have been difficult to target using organic chemistry methods, thus highlighting the potential of employing aliphatic halogenases for the late-stage functionalization of complex natural products. These structurally unique and selectively active natural products are desirable targets as they have already demonstrated their extraordinary power as shuttles to new biological target spaces[71–74].

Future efforts to understand the underlying structural factors to selectively derivatize non-native substrates will help to generalize evolution strategies for this enzyme family and algorithm-driven engineering as well as homology model-based docking approaches will play an important role in accelerating this process. Looking forward, aliphatic halogenases are rapidly becoming an interesting new tool for the development of biologically active molecules to be used, for example, in medicinal and agrochemistry.

Methods

Materials. All chemicals and solvents were purchased from commercial suppliers (Sigma Aldrich, VWR, and Carl Roth) and were used without further purification. Phusion High-Fidelity DNA polymerase, T4 DNA ligase, and all restriction enzymes used in this study were purchased from New England BioLabs (Massachusetts, USA). Gene synthesis was performed by Twist Bioscience (California, USA). Oligonucleotides and sequencing service was provided by Microsynth AG (Balgach, Switzerland).

Initial halogenase panel and protein expression. Genes encoding halogenases in pET28b(+) were purchased from Twist Bioscience. Each plasmid was transformed into *E.coli* BL21(DE3) and the cells were plated on an LB agar plate containing 50 µg/mL kanamycin. A single colony of freshly transformed cells was cultured overnight in 1 mL of LB medium containing 50 µg/mL kanamycin. About 0.1 mL of the culture was used to inoculate 0.9 mL of TB medium supplemented with 50 µg/mL kanamycin and 0.2 mM IPTG (for Fl-Hal) or of 0.9 mL Zymo5052 autoinduction medium[75] supplemented with 50 µg/mL kanamycin (for αKGH) in a 96-well deep-well plate. Expression was carried out for 24 h at 20 °C, 300 rpm (5-cm shaking diameter) using a Duetz system (Kühner AG, Basel, Switzerland). The cells were pelleted by centrifugation at 4000 × *g*, 4 °C, for 15 min, and the supernatant was discarded. The cell pellet was stored in a −80 °C freezer prior to biotransformation reactions.

Combinatorial library WelO5*. WelO5* was subjected to simultaneous saturation mutagenesis of the three hot spots Val81, Ala88, and Ile161, leading to a theoretical library size of $20^3 = 8000$ mutants. The variants were obtained as a pooled gene fragment library from Twist Bioscience (California, USA) and subcloned without a His-tag into a modified pET28b(+) expression vector, in which the nucleotide sequence between the NcoI and NdeI restriction sites was removed and the NcoI replaced by the NdeI restriction site. Consequently, inserting the gene with a terminal stop codon between the NdeI and XhoI restriction sites yields an ORF without His-tag. The cloning was realized with the In-Fusion HD Cloning Plus kit (Takara Bio, Shiga, Japan). The library was amplified with forward primer 5′-AAGGAGATATACATATGTCGAACAACACCATCTCGAC-3′ and reverse primer 5′-GGTGGTGGTGCTCGAGTTAGCTCCAATAGTAGATTTTGTTG-3′ using the DNA polymerase and a standard PCR protocol provided by the kit manufacturer. The gel-purified PCR product (NucleoSpin Gel and PCR Clean-up, Macherey-Nagel, Düren, Germany) was inserted into NdeI/XhoI-linearized pET28b(+) vector (modified) using the In-Fusion enzyme mix. The resulting reaction mixture was utilized to transform competent *E. coli* Stellar™ cells from the kit. After reconstitution in 1 mL SOC medium, 20–50 µL were spread on an LB kanamycin agar plate for transformant count and the remaining cell solution was inoculated into 50 mL LB kanamycin overnight growth at 37 °C. Plasmid isolated from 10 mL culture was used to transform competent *E. coli* BL21(DE3) cells. Clones from LB kanamycin agar plates were sampled for colony PCR to verify the presence of insert prior to sequencing. More than 1000 colonies were picked and grown separately in 96-deep-well plates for DNA Sanger sequencing (Microsynth AG, Balgach, Switzerland). For screening, strains containing empty vector, wild type WelO5*, and other WelO5* variants (positive controls) were included on each plate.

Biotransformation αKGH. The cell pellets were subjected to chemical lysis using 100 µL of 50 mM sodium phosphate buffer (pH 8.0) supplemented with 1 mg/mL lysozyme, 0.5 mg/mL polymyxin B, and 0.01 mg/mL DNase. Incubation was carried out for a minimum of 30 min at 20 °C on a shaking incubator at 850 rpm. Biotransformations were initiated by the addition of 100 µL of sodium phosphate buffer (pH 8.0) containing 2 mM substrate, 220 mM α-ketoglutaric acid sodium salt, 212 mM sodium ascorbate, 1000 mM NaCl, and 2.6 mM ammonium iron(II) sulfate to each well. Assay plates were sealed with breathable membranes and incubated overnight at 20 °C on a shaking incubator at 850 rpm. The reaction was quenched by the addition of 800 µL methanol/water 5:3 mixture to each well and sealed with microplate foil. The plates were shaken at 850 rpm for 30 min prior to centrifugation at 4000 × *g*, 10 °C, for 15 min. After centrifugation, the supernatant was analyzed via LC-MS. The biotransformations were carried out once including the appropriate controls. Predicted variants (selectivity; activity) and best-performing variants (SLP; WVS) were analyzed in triplicates as individual experiments.

Biotransformation Fl-Hal. The cell pellets were subjected to chemical lysis using 100 µL of 25 mM HEPES buffer (pH 7.5) supplemented with 1 mg/mL lysozyme, 0.5 mg/mL polymyxin B, 0.01 mg/mL DNase and incubation for a minimum of 30 min at 20 °C on a shaking incubator at 850 rpm. Biotransformations were initiated by the addition of 100 µL of HEPES buffer (pH 7.5) containing 2 mM substrate, 0.2 mM FAD/FMN, 2 mM NADH/NADPH, 600 mM NaCl, 40 mM Glucose, and 2 µM GDH/Ec-Fre[76]. Incubation and work-up was performed in analogy to the αKGH protocol. The biotransformations with the Fl-Hal library were carried out once including the appropriate controls.

Preparative scale biotransformation. WelO5* SLP variant was used to prepare compound **1a** and **1b** and WelO5* WVS was used to prepare compound **1c**. For the preparation of compound **2a** the variant WelO5* VAA was used. Twenty grams of WelO5* variant cells were resuspended in 100 mL of lysis buffer (50 mM sodium phosphate, pH 8.0) containing 1 mg/mL lysozyme, 0.5 mg/mL polymyxin B, and 0.01 mg/mL DNase in a 2000 mL baffled flask. The cell suspension was shaken for a minimum of 30 min at 20 °C. Reaction was initiated by the addition of 100 mL sodium phosphate buffer (pH 8.0) containing 2 mM substrate, 220 mM α-ketoglutaric acid sodium salt, 212 mM sodium ascorbate, 1000 mM NaCl, and 2.6 mM ammonium iron(II) sulfate. The flask was incubated overnight at 20 °C on a shaking incubator at 100 rpm. About 200 mL methanol were added to the reaction mixture, and the flask was shaken vigorously. The reaction mixture was transferred to a centrifuge bottle and spun down at 4000 × *g* for 15 min. The supernatant was transferred in a round bottom flask, and methanol was removed by a rotary evaporator. The substrate and derivatives were extracted by ethyl acetate (2 × 400 mL), and the organic layer was washed with saturated NaCl solution. The organic layer was combined and dried over sodium sulfate. The solvent was removed by a rotary evaporator to yield a yellowish-brown oil.

LC-MS analysis. Each biotransformation sample was analyzed by LC-MS system (OpenLAB CDS 2.4). The supernatant was injected into an Agilent 1260 HPLC system equipped with a single quadrupole MSD over an Agilent Poroshell 120 EC-C18 column (2.7 µm 2.1 × 50 mm) heated at 40 °C, using water/acetonitrile 95:5 and acetonitrile containing 0.2% formic acid as solvent A and B, respectively. The following LC method was used: 0–1 min, B = 40%; 1–3 min, B = 40−100%; 3–4 min, B = 100%; 4–5 min, B = 100−40%. Fold increase in total chlorination of individual variants was normalized to a parent variant (WelO5* GAP) included as a control.

Construction, expression, and purification of His-tagged WelO5* variants. His-tagged WelO5* enzyme variants (His-wt, His-GAP, His-SLP, His-VLA, and His-WVS) were created to carry out in vitro biocatalysis reactions. The mutated gene fragment encoding each variant was amplified using a primer pair 5′-GTGAGCGGATAACAATTCCCCTCTAG-3′ (forward) and 5′-GCTTTGTTAGCAGCCGGATCTCAG-3′ (reverse) and digested by NdeI and XhoI, which was then ligated into a pET28b(+) vector digested with the same restriction enzymes. The DNA sequence was confirmed by the DNA sequencing service provided by Microsynth AG.

Each plasmid was transformed into *E. coli* BL21(DE3) and the cells were plated on an LB agar plate containing 50 µg/mL kanamycin. A single colony of freshly transformed cells was cultured overnight in 5 mL of LB medium containing 50 µg/mL kanamycin. The culture was used to inoculate 500 mL of TB medium supplemented with 50 µg/mL kanamycin in a baffled Erlenmeyer flask. To monitor the growth of the cells OD_{600} was measured and at an OD_{600} of 0.6–1.0 the culture was induced with IPTG stock solution (final concentration IPTG 100 µM). Expression was carried out for 24 h at 20 °C using 120 rpm (5-cm shaking diameter). The cells were pelleted by centrifugation at 4000 × *g*, 4 °C, for 15 min, and the supernatant was discarded. The cell pellet was stored in a −20 °C freezer prior to purification. Cell pellets were resuspended in 30 mL of protein lysis buffer (50 mM Tris-HCl, pH = 7.4, 500 mM NaCl, 20 mM imidazole, 10 mM β-mercaptoethanol (β-ME), and 0.1% Tween-20) and sonicated over two rounds for 2 min with 1 s intervals on ice and then centrifuged for 30 min at 8000 × *g* at 4 °C. The column (HisTrap™crude; 5 mL, GE Healthcare, Massachusetts, USA) was equilibrated using at least five column volumes of protein lysis buffer. The supernatant was filtered through a 0.45-µm filter and loaded onto the column. After reaching a stable UV baseline the concentration of elution buffer (50 mM Tris, pH = 7.4, 500 mM NaCl, 100 and 250 mM imidazole, and 10 mM β-ME) was raised to 100% to elute the His-tagged protein. The fractions were combined according to the UV spectra (280 nm) and the buffer was exchanged to a buffer containing 50 mM sodium phosphate, pH 8.0. Purified protein was analyzed by SDS-PAGE to ensure its purity. The protein was concentrated using ultra centrifugal filters (Amicon® Ultra 4, cut off 10–30 kDa, Merck Millipore, MA), then flash-frozen using liquid nitrogen and stored at −80 °C. The protein concentration was determined by measuring the protein absorption via a NanoDrop spectrometer (Thermo Fisher Scientific) at 280 nm applying the estimated extinction coefficient of the protein variants (28,880 $M^{-1}cm^{-1}$ for His-GAP, His-SLP, His-VLA and 34,380 $M^{-1}cm^{-1}$ for His-WVS).

In vitro activity assay. In vitro activity assays were carried out in 200 µL of 50 mM sodium phosphate pH 8.0, containing 50 µM purified enzyme, 1 mM substrate, 110 mM α-ketoglutaric acid sodium salt, 106 mM sodium ascorbate, 500 mM sodium salts (NaF, NaCl, NaBr, NaI, NaN_3, and $NaNO_2$), and 1.0 mM ammonium iron(II) sulfate. Ninety-six well plates were sealed with breathable membranes and incubated overnight at 20 °C on a shaking incubator at 850 rpm. The reaction mixtures were quenched with an 800 µL methanol/water mixture (62% methanol). The plate was sealed with microplate foil and shaken at 850 rpm for 30 min prior to centrifugation at 4000 × *g*, 10 °C, for 15 min. Each biotransformation sample was analyzed by LC-MS system using selected ion monitoring (SIM).

Formation of **1a** was quantified through a calibration curve (Supplementary Fig. 11) prepared from known concentrations of the product isolated by the preparative scale biotransformation. As internal standard (ISTD) 0.4 mg/L soraphen C was used.

To determine k_{cat} (chlorination of soraphen A), assays were carried out in an identical manner as the assay described above except that reactions were performed at different substrate concentrations (Supplementary Table 6) and with the addition of 3.8% dimethylformamide (µL/µL). At indicated time points (1, 2, 3, 4, and 5 min), 20 µL of reaction mixture was transferred into 980 µL of methanol/water mixture (methanol:water = 1:1 + 0.4 mg/L soraphen C as ISTD) to quench the reaction. The product formation was monitored by LC-MS and was plotted over time, which was then fitted by linear regression using Microsoft Excel. The observed initial rates were fitted to a substrate inhibition model (Supplementary Eq. 1 and Supplementary Fig. 12) using GraphPad Prism 8.4.0 (nonlinear regression) with the following restraints: $K_m > 0$, $K_i > K_m$. TTN was determined at a substrate concentration of 60 µM and the reaction was quenched at stable product concentration using the same procedure as above. The following enzyme variant concentrations were used: GAP = 5 µM, SLP = 0.5 µM, and VLA = 0.1 and 0.5 µM.

Ligand docking and homology modeling of WelO5* variants. Models of the wild type WelO5* and the WelO5* variants were created using the SWISS-MODEL[56] online server with default parameters. The crystal structure of wild type WelO5 (PDB ID: 5J4R) served as a template for the homology modeling. The docking process was performed using default parameters of Chimera AutoDock Vina[77] and the region of interest was set to default, as this docking is flexible. Each docking result was visually inspected using PyMOL 2.4.1 software.

Machine learning. The label vector was defined as activity or selectivity. The activity label (*A*) was calculated using the formula *A* = *tot. Cl conversion WelO5* mutant* / *tot. Cl conversion WelO5* GAP* whereas *tot. Cl conversion* = (SIM_{1a} + SIM_{1b}) / (SIM_{1a} + SIM_{1b} + SIM_{1c} + SIM_1). The selectivity label (*S*) was calculated using the formula $S = (SIM_{1a} - SIM_{1b}) / (SIM_{1a} + SIM_{1b})$. Amino acids were represented as a 17-dimensional vector, which was obtained by concatenating the five-dimensional T-scale descriptor[63] and additional information about amino acid characteristics[64]. We then defined the feature vector of a sequence by joining the vector representation of its individual amino acids at sites V81X, A88X, I161X, and aggregated them into the 504 × 51-dimensional training matrix. This was used to train a machine learning model, based on the Algorithm 2.1 of Gaussian Processes for Machine Learning (GPML) by Rasmussen and Williams[78], implemented in the scikit-learn python module. We took a similar approach for predictions of selectivity; however, we excluded variants below a peak area threshold and relied on the random forest implementation in scikit -learn for predictions, using the same input features as for activity. To avoid overfitting and to better gauge the generalizability of our model, we cross-validated over ten splits, and model performance was evaluated on the coefficient of determination (R^2), a standard metric for regression problems, achieving an out of fold score of 0.745/0.31 for activity/selectivity respectively (compare predicted vs. measured Supplementary Fig. 13). Inference occurred on the remaining sequence space, which was preprocessed exactly like the training data, at every fold during cross-validation. The code, data, and supplementary information, such as amino acid encodings, can be accessed at: [https://github.com/ccbiozhaw/MLevo].

Reporting summary. Further information on research design is available in the Nature Research Reporting Summary linked to this article.

Data availability

Source data are provided with this paper. WelO5 crystal structure used as template for SWISS-MODEL homology modeling can be accessed via PDB ID: 5J4R. The authors declare that all the data supporting the findings of this work are available within the article and its Supplementary Information and the provided Source Data. Source data are provided with this paper.

Code availability

Training data and scripts used to predict enzyme function are available at https://github.com/ccbiozhaw/MLevo, https://doi.org/10.5281/zenodo.5665270

Received: 10 August 2021; Accepted: 17 December 2021;
Published online: 18 January 2022

References

1. Harris, C. M., Kannan, R., Kopecka, H. & Harris, T. M. The role of the chlorine substituents in the antibiotic vancomycin: preparation and characterization of mono- and didechlorovancomycin. *J. Am. Chem. Soc.* **107**, 6652–6658 (1985).
2. Schönherr, H. & Cernak, T. Profound methyl effects in drug discovery and a call for new C-H methylation reactions. *Angew. Chem. Int. Ed.* **52**, 12256–12267 (2013).
3. Cernak, T., Dykstra, K. D., Tyagarajan, S., Vachal, P. & Krska, S. W. The medicinal chemist's toolbox for late stage functionalization of drug-like molecules. *Chem. Soc. Rev.* **45**, 546–576 (2016).
4. Hong, B., Luo, T. & Lei, X. Late-stage diversification of natural products. *ACS Cent. Sci.* **6**, 622–635 (2020).
5. Börgel, J. & Ritter, T. Late-stage functionalization. *Chem.* **6**, 1877–1887 (2020).
6. Weichold, V., Milbredt, D. & Van Pée, K. H. Specific enzymatic halogenation - from the discovery of halogenated enzymes to their applications in vitro and in vivo. *Angew. Chem. Int. Ed.* **55**, 6374–6389 (2016).
7. Kelly, C. B. & Padilla-Salinas, R. Late stage C-H functionalization: via chalcogen and pnictogen salts. *Chem. Sci.* **11**, 10047–10060 (2020).
8. Petrone, D. A., Ye, J. & Lautens, M. Modern transition-metal-catalyzed carbon-halogen bond formation. *Chem. Rev.* **116**, 8003–8104 (2016).
9. Hartwig, J. F. & Larsen, M. A. Undirected, homogeneous C-H bond functionalization: challenges and opportunities. *ACS Cent. Sci.* **2**, 281–292 (2016).
10. Latham, J., Brandenburger, E., Shepherd, S. A., Menon, B. R. K. & Micklefield, J. Development of halogenase enzymes for use in synthesis. *Chem. Rev.* **118**, 232–269 (2018).
11. Büchler, J., Papadopoulou, A. & Buller, R. Recent advances in flavin-dependent halogenase biocatalysis: sourcing, engineering, and application. *Catalysts* **9**, 1030 (2019).
12. Voss, M., Honda Malca, S. & Buller, R. Exploring the biocatalytic potential of Fe/α-ketoglutarate-dependent halogenases. *Chem. Eur. J.* **26**, 7336–7345 (2020).
13. Wu, S., Snajdrova, R., Moore, J. C., Baldenius, K. & Bornscheuer, U. T. Biocatalysis: enzymatic synthesis for industrial applications. *Angew. Chem. Int. Ed.* **60**, 88–119 (2021).
14. Gkotsi, D. S., Dhaliwal, J., McLachlan, M. M., Mulholand, K. R. & Goss, R. J. Halogenases: powerful tools for biocatalysis (mechanisms applications and scope). *Curr. Opin. Chem. Biol.* **43**, 119–126 (2018).
15. Galonić, D. P., Vaillancourt, F. H. & Walsh, C. T. Halogenation of unactivated carbon centers in natural product biosynthesis: trichlorination of leucine during barbamide biosynthesis. *J. Am. Chem. Soc.* **128**, 3900–3901 (2006).
16. Vaillancourt, F. H., Yin, J. & Walsh, C. T. SyrB2 in syringomycin E biosynthesis is a nonheme Fe^{II} α-ketoglutarate- and O_2-dependent halogenase. *Proc. Natl Acad. Sci. USA* **102**, 10111–10116 (2005).
17. Ueki, M. et al. Enzymatic generation of the antimetabolite γ, γ-dichloroaminobutyrate by NRPS and mononuclear iron halogenase action in a Streptomycete. *Chem. Biol.* **13**, 1183–1191 (2006).
18. Vaillancourt, F. H., Yeh, E., Vosburg, D. A., O'Connor, S. E. & Walsh, C. T. Cryptic chlorination by a non-haem iron enzyme during cyclopropyl amino acid biosynthesis. *Nature* **436**, 1191–1194 (2005).
19. Pratter, S. M. et al. More than just a halogenase: modification of fatty acyl moieties by a trifunctional metal enzyme. *ChemBioChem* **15**, 567–574 (2014).
20. Khare, D. et al. Conformational switch triggered by α-ketoglutarate in a halogenase of curacin a biosynthesis. *Proc. Natl Acad. Sci. USA* **107**, 14099–14104 (2010).
21. Hillwig, M. L. & Liu, X. A new family of iron-dependent halogenases acts on freestanding substrates. *Nat. Chem. Biol.* **10**, 921–923 (2014).
22. Zhu, Q. & Liu, X. Characterization of non-heme iron aliphatic halogenase WelO5* from *Hapalosiphon welwitschii* IC-52-3: Identification of a minimal protein sequence motif that confers enzymatic chlorination specificity in the biosynthesis of welwitindolelinones. *Beilstein. J. Org. Chem.* **13**, 1168–1173 (2017).
23. Duewel, S. et al. Directed evolution of an Fe^{II}-dependent halogenase for asymmetric C(sp^3)-H chlorination. *ACS Catal.* **10**, 1272–1277 (2020).

24. Hillwig, M. L., Zhu, Q., Ittiamornkul, K. & Liu, X. Discovery of a promiscuous non-heme iron halogenase in ambiguine alkaloid biogenesis: implication for an evolvable enzyme family for late-stage halogenation of aliphatic carbons in small molecules. *Angew. Chem. Int. Ed.* **55**, 5780–5784 (2016).
25. Neugebauer, M. E. et al. A family of radical halogenases for the engineering of amino-acid-based products. *Nat. Chem. Biol.* **15**, 1009–1016 (2019).
26. Kim, C. Y. et al. The chloroalkaloid (−)-acutumine is biosynthesized via a Fe(II)- and 2-oxoglutarate-dependent halogenase in Menispermaceae plants. *Nat. Commun.* **11**, 1867 (2020).
27. Zhao, C. et al. An Fe^{2+}- and α-ketoglutarate-dependent halogenase acts on nucleotide substrates. *Angew. Chem. Int. Ed.* **59**, 9478–9484 (2020).
28. Brown, S. & O'Connor, S. E. Halogenase engineering for the generation of new natural product analogues. *ChemBioChem* **16**, 2129–2135 (2015).
29. Payne, J. T., Andorfer, M. C. & Lewis, J. C. Engineering flavin-dependent halogenases. *Methods Enzymol.* **575**, 93–126 (2016).
30. van Pée, K. H., Milbredt, D., Patallo, E. P., Weichold, V. & Gajewi, M. Application and modification of flavin-dependent halogenases. *Methods Enzymol.* **575**, 65–92 (2016).
31. Shepherd, S. A. et al. Extending the biocatalytic scope of regiocomplementary flavin-dependent halogenase enzymes. *Chem. Sci.* **6**, 3454–3460 (2015).
32. Shepherd, S. A. et al. A structure-guided switch in the regioselectivity of a tryptophan halogenase. *ChemBioChem* **17**, 821–824 (2016).
33. Minges, H. et al. Targeted enzyme engineering unveiled unexpected patterns of halogenase stabilization. *ChemCatChem* **12**, 818–831 (2020).
34. Yamada, R., Higo, T., Yoshikawa, C., China, H. & Ogino, H. Improvement of the stability and activity of the BPO-A1 haloperoxidase from *Streptomyces aureofaciens* by directed evolution. *J. Biotechnol.* **192**, 248–254 (2014).
35. Yamada, R. et al. Random mutagenesis and selection of organic solvent-stable haloperoxidase from *Streptomyces aureofaciens*. *Biotechnol. Prog.* **31**, 917–924 (2015).
36. Roy, A. D., Grüschow, S., Cairns, N. & Goss, R. J. M. Gene expression enabling synthetic diversification of natural products: chemogenetic generation of pacidamycin analogs. *J. Am. Chem. Soc.* **132**, 12243–12245 (2010).
37. Runguphan, W. & O'Connor, S. E. Diversification of monoterpene indole alkaloid analogs through cross-coupling. *Org. Lett.* **15**, 2850–2853 (2013).
38. Durak, L. J., Payne, J. T. & Lewis, J. C. Late-stage diversification of biologically active molecules via chemoenzymatic C-H functionalization. *ACS Catal.* **6**, 1451–1454 (2016).
39. Latham, J. et al. Integrated catalysis opens new arylation pathways via regiodivergent enzymatic C-H activation. *Nat. Commun.* **7**, 11873 (2016).
40. Gkotsi, D. S. et al. A marine viral halogenase that iodinates diverse substrates. *Nat. Chem.* **11**, 1091–1097 (2019).
41. Andorfer, M. C., Park, H. J., Vergara-Coll, J. & Lewis, J. C. Directed evolution of RebH for catalyst-controlled halogenation of indole C-H bonds. *Chem. Sci.* **7**, 3720–3729 (2016).
42. Payne, J. T., Poor, C. B. & Lewis, J. C. Directed evolution of RebH for site-selective halogenation of large biologically active molecules. *Angew. Chem. Int. Ed.* **54**, 4226–4230 (2015).
43. Hayashi, T. et al. Evolved aliphatic halogenases enable regiocomplementary C−H functionalization of a pharmaceutically relevant compound. *Angew. Chem. Int. Ed.* **58**, 18535–18539 (2019).
44. Naini, A., Sasse, F. & Brönstrup, M. The intriguing chemistry and biology of soraphens. *Nat. Prod. Rep.* **36**, 1394–1411 (2019).
45. Shen, Y., Volrath, S. L., Weatherly, S. C., Elich, T. D. & Tong, L. A mechanism for the potent inhibition of eukaryotic acetyl-coenzyme A carboxylase by soraphen A, a macrocyclic polyketide natural product. *Mol. Cell* **16**, 881–891 (2004).
46. Wei, J. & Tong, L. Crystal structure of the 500-kDa yeast acetyl-CoA carboxylase holoenzyme dimer. *Nature* **526**, 723–727 (2015).
47. Weissman, K. J. & Müller, R. Myxobacterial secondary metabolites: bioactivities and modes-of-action. *Nat. Prod. Rep.* **27**, 1276–1295 (2010).
48. Canterbury, D. P. et al. Synthesis of C11-desmethoxy soraphen A$_{1\alpha}$: a natural product analogue that inhibits acetyl-CoA carboxylase. *ACS Med. Chem. Lett.* **4**, 1244–1248 (2013).
49. Hill, A. M. & Thompson, B. L. Novel soraphens from precursor directed biosynthesis. *Chem. Commun.* **3**, 1358–1359 (2003).
50. Taylor, R. E., Chen, Y., Galvin, G. M. & Pabba, P. K. Conformation-activity relationships in polyketide natural products. Towards the biologically active conformation of epothilone. *Org. Biomol. Chem.* **2**, 127–132 (2004).
51. Zirkle, R., Ligon, J. M. & Molnár, I. Heterologous production of the antifugal polyketide antibiotic soraphen A of Sorangium cellulosum So ce26 in *Streptomyces lividans*. *Microbiology* **150**, 2761–2774 (2004).
52. Raymer, B. et al. Synthesis and characterization of a BODIPY-labeled derivative of soraphen A that binds to acetyl-CoA carboxylase. *Bioorganic Med. Chem. Lett* **19**, 2804–2807 (2009).
53. Bedorf, N. et al. Mikrobiologisches Verfahren zur Herstellung agrarchemisch verwendbarer mikrobizider makrozyklischer Lactonderivate. EP 358606 A2 (1990).
54. Poor, C. B., Andorfer, M. C. & Lewis, J. C. Improving the stability and catalyst lifetime of the halogenase RebH by directed evolution. *ChemBioChem* **15**, 1286–1289 (2014).
55. Matthews, M. L. et al. Substrate positioning controls the partition between halogenation and hydroxylation in the aliphatic halogenase, SyrB2. *Proc. Natl Acad. Sci. USA* **106**, 17723–17728 (2009).
56. Waterhouse, A. et al. SWISS-MODEL: homology modelling of protein structures and complexes. *Nucleic Acids Res.* **46**, W296–W303 (2018).
57. Li, A. et al. Beating bias in the directed evolution of proteins: combining high-fidelity on-chip solid-phase gene synthesis with efficient gene assembly for combinatorial library construction. *ChemBioChem* **19**, 196–196 (2018).
58. Banyai, W., Chen, S., Fernandez, A., Indermuhle, P. & Peck, B. J. De novo synthesized gene libraries. Patent WO2015021080A3 (2015).
59. Reetz, M. T., Kahakeaw, D. & Lohmer, R. Addressing the numbers problem in directed evolution. *ChemBioChem* **9**, 1797–1804 (2008).
60. Romero, P. A., Krause, A. & Arnold, F. H. Navigating the protein fitness landscape with gaussian processes. *Proc. Natl Acad. Sci. USA* **110**, E193–E201 (2013).
61. Saito, Y. et al. Machine-learning-guided mutagenesis for directed evolution of fluorescent proteins. *ACS Synth. Biol.* **7**, 2014–2022 (2018).
62. Vornholt, T. et al. Systematic engineering of artificial metalloenzymes for new-to-nature reactions. *Sci. Adv.* **7**, eabe4208 (2021).
63. Tian, F., Zhou, P. & Li, Z. T-scale as a novel vector of topological descriptors for amino acids and its application in QSARs of peptides. *J. Mol. Struct.* **830**, 106–115 (2007).
64. Ibraheem, Z. O., Abd Majid, R., Noor, S. M., Sedik, H. M. & Basir, R. Role of different pfcrt and pfmdr-1 mutations in conferring resistance to antimalaria drugs in *Plasmodium falciparum*. *Malar. Res. Treat.* **2014**, 950424 https://doi.org/10.1155/2014/950424 (2014).
65. Lin, Y. et al. Substrate inhibition kinetics for cytochrome P450-catalyzed reactions. *Drug Metab. Dispos.* **29**, 368–374 (2001).
66. Mitchell, A. J. et al. Structural basis for halogenation by iron-and 2-oxo-glutarate-dependent enzyme WelO5. *Nat. Chem. Biol.* **12**, 636–640 (2016).
67. Matthews, M. L. et al. Direct nitration and azidation of aliphatic carbons by an iron-dependent halogenase. *Nat. Chem. Biol.* **10**, 209–215 (2014).
68. The European Green Deal. https://ec.europa.eu/info/sites/info/files/europea (2019).
69. Mazurenko, S., Prokop, Z. & Damborsky, J. Machine learning in enzyme engineering. *ACS Catal.* **10**, 1210–1223 (2020).
70. Siedhoff, N. E., Schwaneberg, U. & Davari, M. D. Machine learning-assisted enzyme engineering. *Methods Enzymol.* **643**, 281–315 (2020).
71. Bade, R., Chan, H. F. & Reynisson, J. Characteristics of known drug space. Natural products, their derivatives and synthetic drugs. *Eur. J. Med. Chem.* **45**, 5646–5652 (2010).
72. Newman, D. J. & Cragg, G. M. Natural products as sources of new drugs over the nearly four decades from 01/1981 to 09/2019. *J. Nat. Prod.* **83**, 770–803 (2020).
73. Atanasov, A. G. et al. Natural products in drug discovery: advances and opportunities. *Nat. Rev. Drug Discov.* **20**, 200–216 (2021).
74. Loiseleur, O. Natural products in the discovery of agrochemicals. *Chimia* **71**, 810–822 (2017).
75. Studier, F. W. Protein production by auto-induction in high density shaking cultures. *Protein Expr. Purif.* **41**, 207–234 (2005).
76. Spyrou, G. et al. Characterization of the flavin reductase gene (fre) of *Escherichia coli* and construction of a plasmid for overproduction of the enzyme. *J. Bacteriol.* **173**, 3673–3679 (1991).
77. Trott, O. & Olson, A. J. AutoDock Vina: improving the speed and accuracy of docking with a new scoring function, efficient optimization, and multithreading. *J. Comput. Chem.* **31**, 455–461 (2010).
78. Rasmussen, C. E. In *Advanced Lectures on Machine Learning. ML 2003* (eds Bousquet, O., von Luxburg, U., & Rätsch, G.) (Springer, 2004).
79. Galonić, D. P., Barr, E. W., Walsh, C. T., Bollinger, J. M. & Krebs, C. Two interconverting Fe(IV) intermediates in aliphatic chlorination by the halogenase CytC3. *Nat. Chem. Biol.* **3**, 113–116 (2007).

Acknowledgements

We thank Myriam Baalouch for her help with the synthesis and analytics of the soraphen compounds, Dirk Balmer for the biological testing of the compounds, Dianne Irwin for the purification of the chlorinated products, Federico Dapiaggi for his help with the docking experiments and helpful discussions, and Leonard Hagmann for his help with the NMR analysis of the products. Furthermore, we thank An Vandemeulebroucke for the helpful discussion of the kinetic results. This work was supported by the Swiss State Secretariat for Education, Research and Innovation (Federal project contributions 2017–2020, P-14: Innovation in Biocatalysis) and was created as part of NCCR Catalysis, a National Centre of Competence in Research funded by the Swiss National Science Foundation (Grant number 180544).

Author contributions

J.B., A.L., C.L.C., O.L. and R.B. designed the research. J.B. carried out most of the experiments and performed the docking analysis. S.H.M. and M.V. constructed the combinatorial enzyme library. D.P. carried out the machine learning predictions. O.A. synthesized the soraphen analogs. C.L.C. analyzed the NMR structures. J.B., N.J.T., U.T.B., C.L.C., O.L. and R.B. discussed the results and wrote the manuscript.

Competing interests

Author O.A. is employed by Idorsia Pharmaceuticals Ltd and authors C.C., A.L. and O.L. are employees of Syngenta Crop Protection AG. The remaining authors declare no competing interests.

Additional information

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41467-022-27999-1.

Correspondence and requests for materials should be addressed to Olivier Loiseleur or Rebecca Buller.

Peer review information *Nature Communications* thanks Kinshuk Raj Srivastava and the other anonymous reviewer(s) for their contribution to the peer review this work. Peer reviewer reports are available.

Reprints and permission information is available at http://www.nature.com/reprints

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this license, visit http://creativecommons.org/licenses/by/4.0/.

© The Author(s) 2022

Supplementary Information

Algorithm-aided engineering of aliphatic halogenase WelO5* for the asymmetric late-stage functionalization of soraphens

Johannes Büchler,[a, b] Sumire Honda Malca[a], David Patsch,[a, c] Moritz Voss,[a] Nicholas J. Turner[b], Uwe T. Bornscheuer[c], Oliver Allemann[d, e], Camille Le Chapelain[d], Alexandre Lumbroso[d], Olivier Loiseleur*[d] and Rebecca Buller*[a]

---

[a] Competence Center for Biocatalysis, Institute of Chemistry and Biotechnology, Zurich University of Applied Sciences, Einsiedlerstrasse 31, 8820 Wädenswil, Switzerland, rebecca.buller@zhaw.ch
[b] School of Chemistry, The University of Manchester, Manchester Institute of Biotechnology, Manchester M1 7DN, United Kingdom
[c] Institute of Biochemistry, Dept. of Biotechnology & Enzyme Catalysis, Greifswald University, Felix-Hausdorff-Strasse 4, 17487 Greifswald, Germany
[d] Syngenta Crop Protection AG, Schaffhauserstrasse 101, 4332 Stein, Switzerland, olivier.loiseleur@syngenta.com
[e] Idorsia Pharmaceuticals Ltd, Hegenheimermattweg 91, 4123 Allschwil, Switzerland

Contents

I. Supplementary Tables 2
II. Supplementary Figures 8
III. Supplementary Methods 23
  Chemical synthesis methods 23
  Synthesis and characterization of soraphen C and soraphen analogues 24
  Purification of bio-extracts 29
  Analytical data of the halogenated and hydroxylated products 30
  NMR spectra 33
  Methods for biological testing and BP80 determination 58
  Combinatorial library WelO5* 65
IV. Supplementary References 66

I. Supplementary Tables

Supplementary Table 1. Fl-Hal panel.

Natural enzymes	
Name	Source organism
Bmp5[1]	*P. luteoviolacea* 2ta16
PyrH[2]	*Streptomyces ru- gosporus* LL-42D005
KtzR[3]	*Kutzneria* sp. 744
Th-Hal[4]	*Streptomyces violaceusniger* SPC6
SttH[5]	*Streptomyces toxytricini* NRRL 15443
PrnA[6]	*Pseudomonas fluorescens*
KtzQ[3]	*Kutzneria* sp. 744
RebH[7]	*Lechevalieria aerocolonigenes* (strain 39243)
ThaL[8]	*Streptomyces albogriseolus*
PrnC[6]	*Pseudomonas fluorescens*
RadH[9]	*Chaetomium chiversii*
MalA’[10]	*Malbranchea graminicola* (086937A)
Rdc2[11]	*Pochonia chlamydosporia*
ChlA[12]	*Dictyostelium discoideum*
Engineered enzymes	
Name	Name
PrnA_F103A[13]	RebH_3SS[14]
PrnA_E450K_F454K[13]	RebH_4V[14]
SttH_Triple	RebH_5LS[15]
RebH_0S[15]	RebH_6TL[15]
RebH_1PVM[14]	RebH_8F[15]
RebH_2T[14]	RebH_10S[15]
RebH_3S[14]	RebH_Thermo[16]

Supplementary Table 2. αKGHs panel.

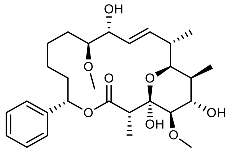
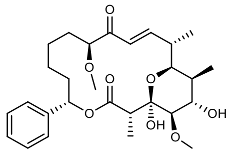
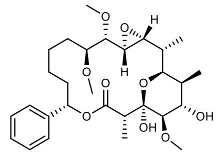
Natural enzymes	
Name	Source organism
WelO5[17]	*Hapalosiphon welwitschii* UTEX B1830
WelO5*[18]	*Hapalosiphon welwitschii* IC-52-3
AmbO5[19]	*Fischerella ambigua* UTEX1903
Engineered enzymes	
Name	Name
WelO5*_N74L	WelO5*_V81L
WelO5*_V81T	WelO5*_I84F
WelO5*_A88G	WelO5*_A88S
WelO5*_A88T	WelO5*_V90P
WelO5*_P153F	WelO5*_P153K
WelO5*_161A	WelO5*_I161D
WelO5*_I161G	WelO5*_I161R
WelO5*_I161S	WelO5*_I161T
WelO5*_I161E	WelO5*_I225M
WelO5*_E76V_V81L	WelO5*_ V81G_I161G
WelO5*_ V81G_I161P	WelO5*_ V81L_A88T
WelO5*_ V81L_I161D	WelO5*_ V81L_I161M
WelO5*_ V81L_I161V	WelO5*_ V81R_I161D
WelO5*_ V81R_I161G	WelO5*_ V81R_I161S
SadA_D157G[20]	

Supplementary Table 3. Ranking of the predicted variants using machine learning. Variants were predicted towards increase in activity and towards increase in selectivity. The produced "activity" variants were chosen on the highest activity predictions. The "selectivity for 1b" variants were chosen on high predicted selectivity towards product **1b** with the additional threshold that the activity predictions had to be higher than 0.6. The activity label (*A*) was calculated using the formula $A = \text{tot. Cl conversion WelO5* variant} / \text{tot. Cl conversion WelO5* GAP}$ (*tot. Cl conversion* = ($SIM_{1a} + SIM_{1b}$) / ($SIM_{1a} + SIM_{1b} + SIM_{1c} + SIM_{1}$)). The selectivity label (*S*) was calculated using the formula $S = (SIM_{1a} – SIM_{1b}) / (SIM_{1a} + SIM_{1b})$.

Activity				**Selectivity 1b**		
Mutant	Ranking activity	Activity p/m[#]	Selectivity p/m[#]	Mutant	Selectivity p/m[#]	Activity p/m[#]
SIP	1	10.2/10.4	0.32/0.08	AHS	-0.77/-0.75	0.7/2.7
VIA	2	9.3/11.4	0.32/0.23	MHS	-0.68/-0.77	1.2/0.9
AIP	3	9.2/11.2	0.07/0.08	AMS	-0.67/*	0.8/*
ALP	4	8.2/11.3	0.56/0.70	VHS	-0.60/-0.75	1.5/1.9
CIA	5	8.0/*	0.25/*	LHT	-0.55/-1.0	0.7/0.5
CIP	6	7.84/*	0.30/*	LHA	-0.53/-0.67	1.8/1.9
SVP	7	7.6/8.5	0.63/0.52	AHG	-0.52/-1.0	1.2/4.2
SIA	8	7.4/5.9	0.24/0.28	AHA	-0.52/*	1.2/*
CLP	9	7.2/*	0.74/*	LMS	-0.51/0.52	0.6/0.6
VLA	10	7.2/11.0	0.77/0.72	LHG	-0.51/-1.0	1.7/3.8

* these variants were not measured, [#]predicted/measured

Supplementary Table 4. Activity of WelO5* variants for the functionalization of soraphen derivatives. The halogenase variants were capable to produce multiple products (as reported). Each of the reported products were derivatized once only as observed by selected ion monitoring.

Compound	Observed products	WelO5* variant
 Soraphen C, **2**	- 2 chlorinated products - 2 hydroxylated products	- VAA[+] - ILV[°]
 **3**	- 4 chlorinated products - 3 hydroxylated products	- SLP[+] - SLP[°]
 **4**	- 5 chlorinated products - 2 hydroxylated products	- SLP[+] - SHP[°]

[+] Variant showing the highest amount of total chlorination; [°] Variant showing the highest amount of total hydroxylation

Supplementary Table 5. Selected examples of the application of machine learning used for the engineering of enzymes.

Year	Enzyme	Target	Number of variants	Perc. Covered /%	Reference
2007	halohydrin dehalogenase	activity	30-150 at each round for 18 rounds	~$(3*N/2^N) * 100$ N= mutation sites	Fox et al.[21]
2012	epoxide hydrolase	enantioselectivity	95	23.8	Feng et al.[22]
2018	green flourescent protein	color change	218	0.14	Saito et al.[23]
2018	epoxide hydrolase	enantioselectivity	37	7.4	Cadet et al.[24]
2019	nitric oxide dioxygenase	stereodivergence	445 over 2 rounds	0.6 - 8.9	Wu et al.[25]
2021	artificial metalloenzymes	activity	400	80.0	Vornholt et al.[26]

Supplementary Table 6. Enzyme and substrate concentrations used for the kinetic experiments.

Variants	Enzyme conc. / µM	Substrate conc. / µM
GAP	20	40, 70, 100, 150, 200, 250,500, 750
VLA	2	40, 70, 100, 150, 200, 250,500, 750
SLP	2	40, 70, 100, 150, 200, 250,500, 750

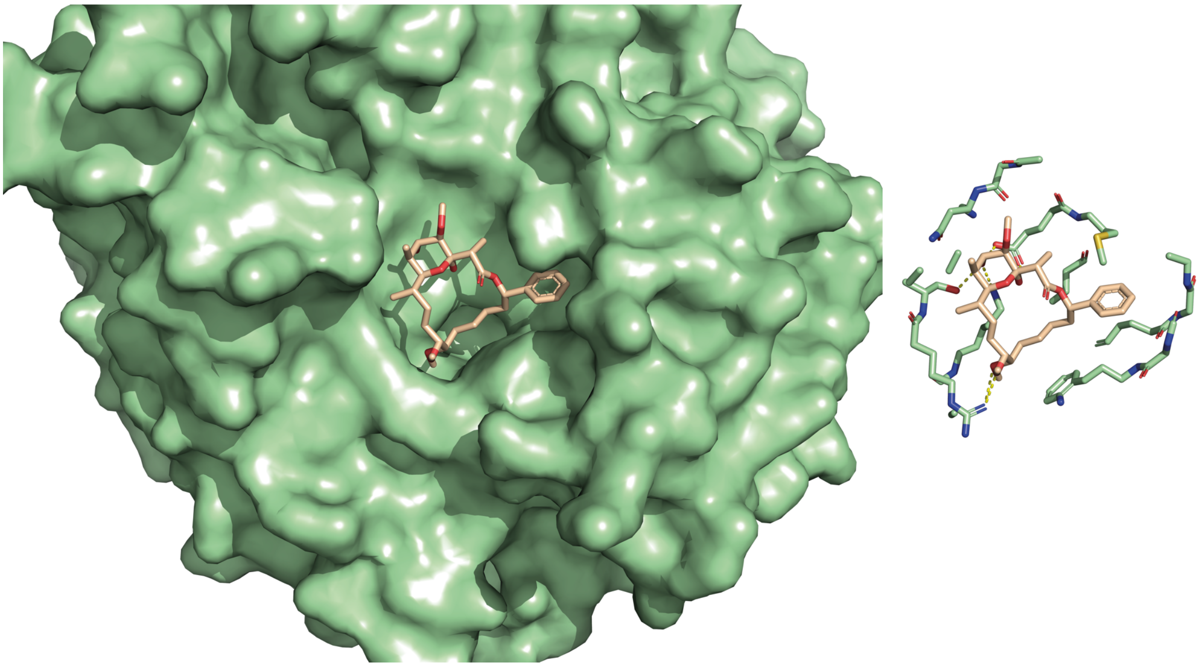
Supplementary Equation 1. Substrate inhibition model.

$$v_0 = \frac{v_{max} * [S]_0}{K_m + [S]_0 + \frac{[S]_0^2}{K_i}}$$

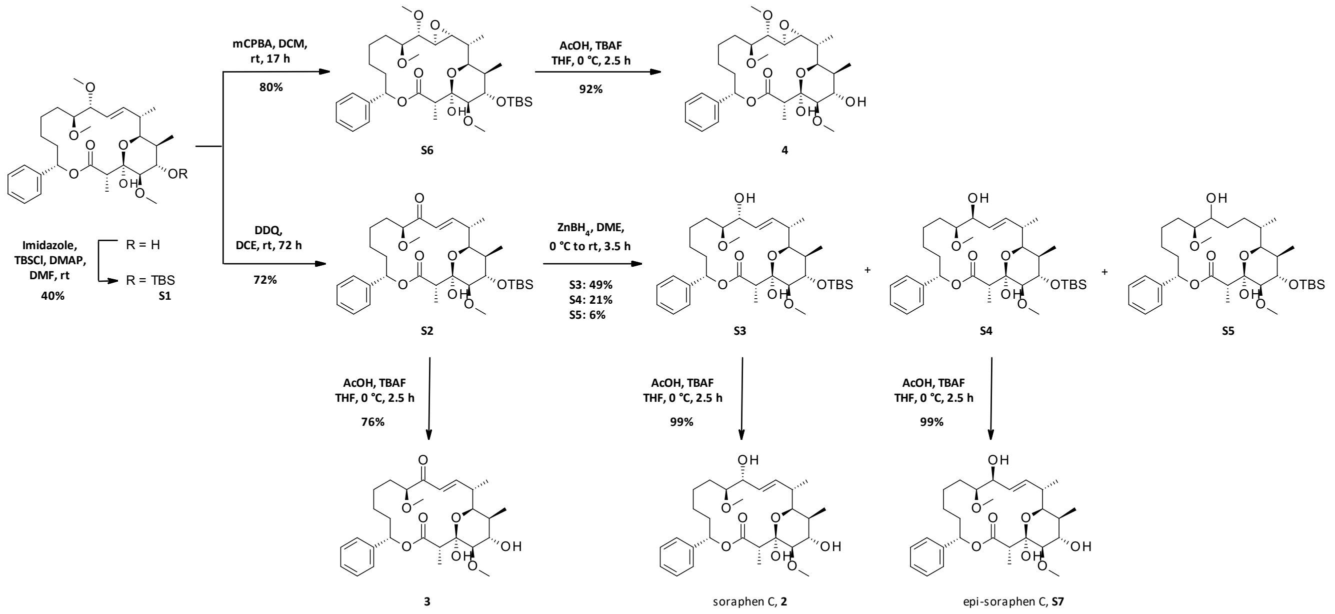
Supplementary Table 7. Comparison of ^{13}C-NMR chemical shifts between soraphen C (reported[2] and synthesized) and synthesized epi-soraphen C.

^{13}C-NMR shifts of fermented soraphen C (δ in ppm)[2]	^{13}C-NMR shifts of synthesized soraphen C (δ in ppm)	Δ1 (ppm)	^{13}C-NMR shifts of synthesized epi-soraphen C (δ in ppm)	Δ2 (ppm)
170.6	170.78	-0.2	172.82	-2.2
141.0	141.12	-0.1	139.69	1.3
137.3	137.45	-0.1	137.96	-0.7
128.6	128.72	-0.1	128.76	-0.2
128.2	128.29	-0.1	128.70	-0.5
126.2	126.35	-0.1	128.46	-2.3
125.0	125.18	-0.2	127.09	-2.1
99.4	99.60	-0.2	99.91	-0.5
83.7	83.89	-0.2	82.86	0.8
76.1	76.28	-0.2	77.57	-1.5
74.9	75.04	-0.1	76.33	-1.4
74.6	74.81	-0.2	73.58	1.0
72.5	72.66	-0.2	71.74	0.8
68.8	68.98	-0.2	68.93	-0.1
57.6	57.78	-0.2	57.49	0.1
57.3	57.47	-0.2	57.45	-0.2
46.2	46.35	-0.1	45.42	0.8
35.8	36.01	-0.2	36.84	-1.0
35.6	35.79	-0.2	35.46	0.1
35.2	35.32	-0.1	35.02	0.2
29.4	29.57	-0.2	27.58	1.8
26.0	26.14	-0.1	25.12	0.9
23.0	23.17	-0.2	22.57	0.4
12.5	12.64	-0.1	16.57	-4.1
11.7	11.83	-0.1	12.55	-0.9
10.3	10.49	-0.2	10.55	-0.3

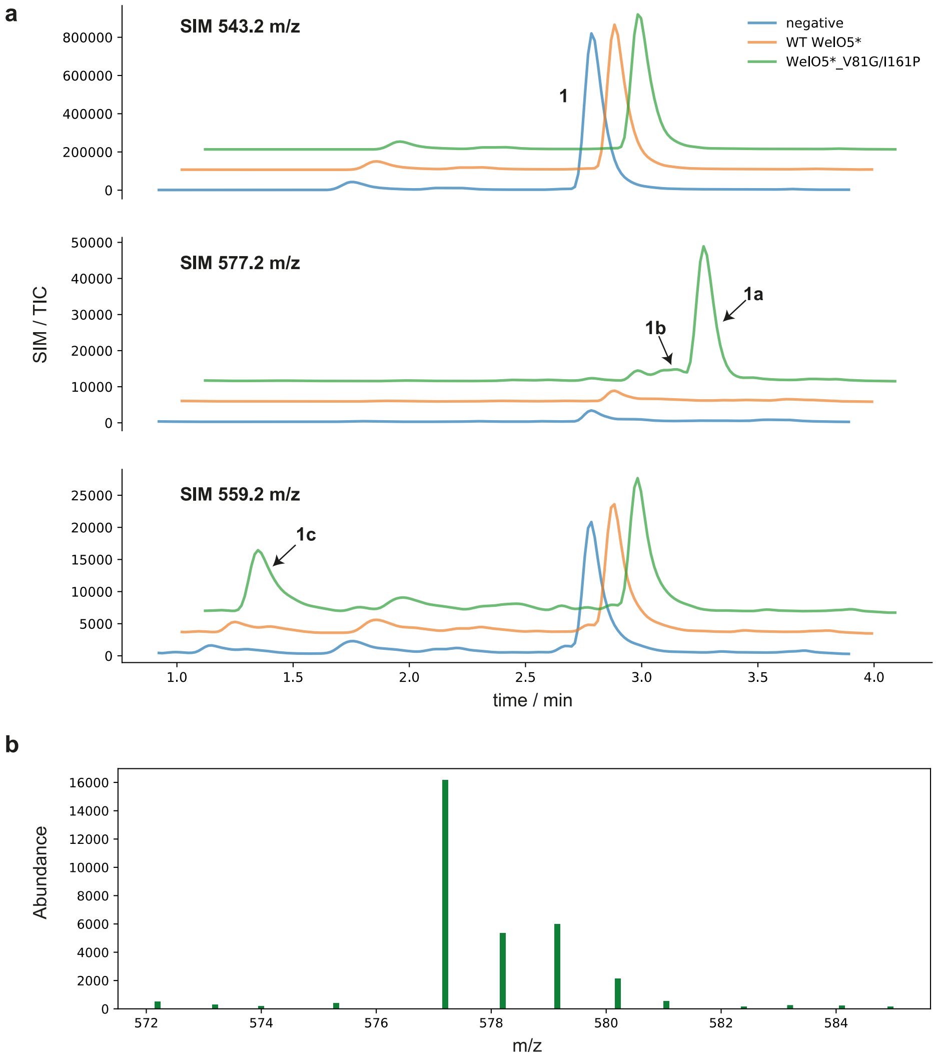
II. Supplementary Figures



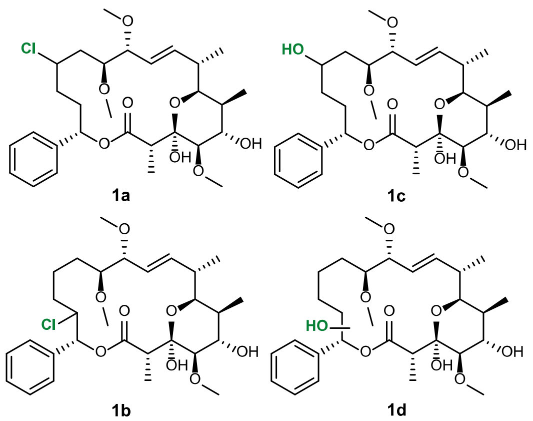
Supplementary Figure 1. Crystal structure of soraphen A bound to the BC domain of yeast acetyl-coenzyme A carboxylases. The crystal structure (green) reveals the active conformation of the macrocycle (PDB ID: 1W96). Right: Visualization of the interactions between the soraphen A (wheat sticks) and the residues of the BC domain (green sticks).



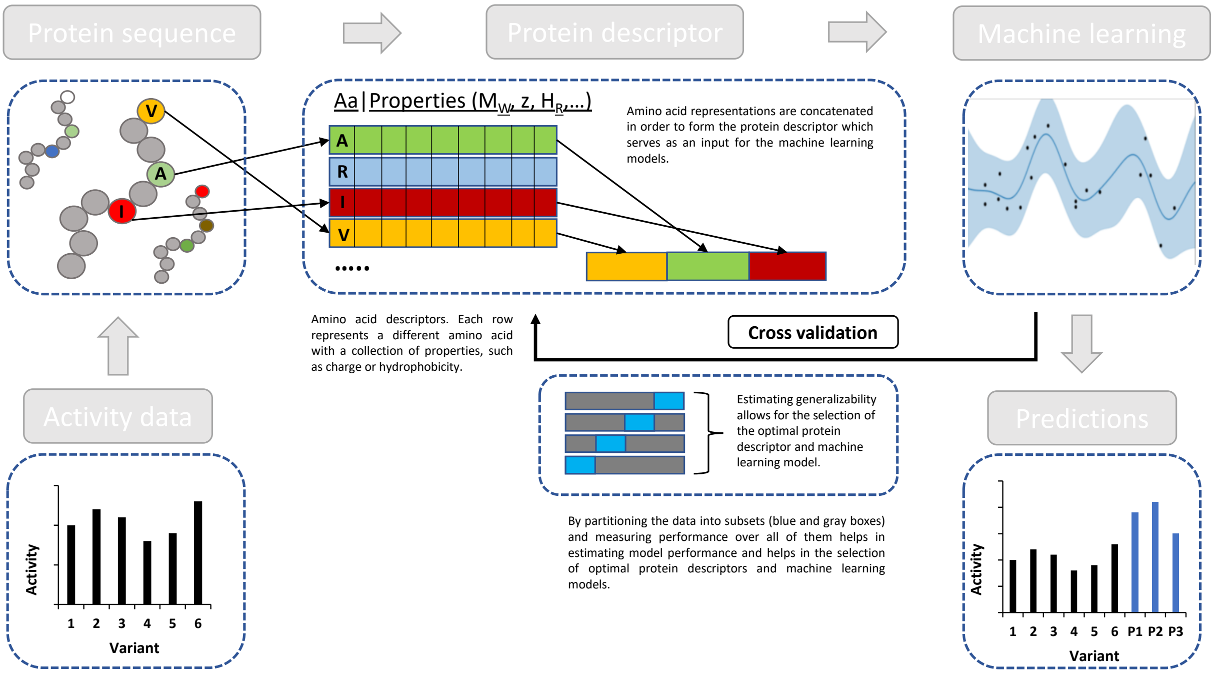
Supplementary Figure 2. Synthesis scheme to obtain soraphen C and soraphen analogues. Individual reaction steps are described in the Supplementary Methods.



Supplementary Figure 3. LC-MS analysis of the biotransformation of soraphen A. a Selected ion chromatograms (SIM) of the m/z values of interest. Biotransformation using negative control (blue), WT WelO5* (orange) and WelO5*_V81G_I161P (green) are compared. The top chromatogram shows the trace of soraphen A (543.2 m/z = **1**+Na+H^+), the middle chromatogram shows two species corresponding to chlorinated soraphen A (577.2 m/z = **1**+Na+H^++Cl^{35}) and the bottom chromatogram shows one hydroxylated soraphen A species (577.2 m/z = **1**+Na+H^++OH). **b** MS chart of the chlorinated species showing the characteristic M: M + 2 = 3 : 1 isotopic pattern of a chlorinated compound.



Supplementary Figure 4. Biotransformation products of soraphen A. Observed products of the biotransformation reactions of soraphen A with WelO5* variants.



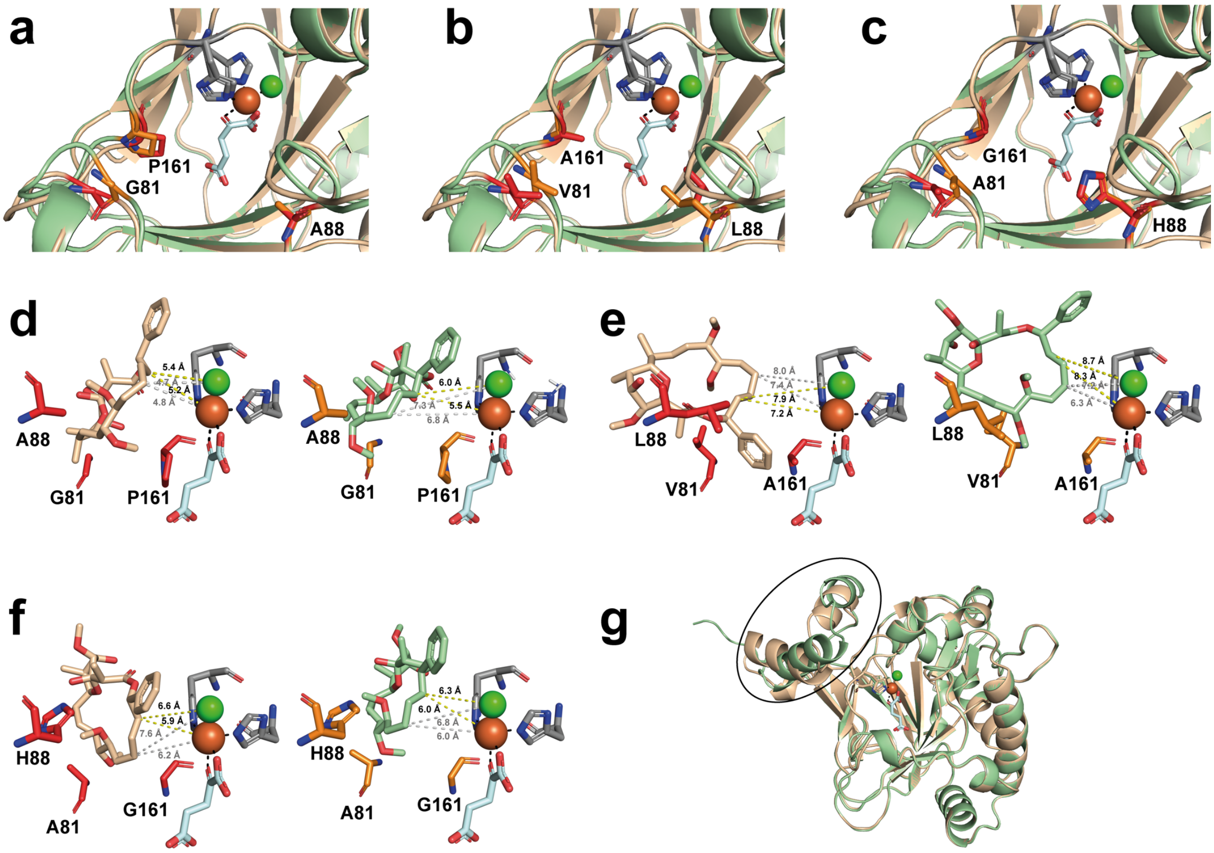
Amino acid representations are concatenated in order to form the protein descriptor which serves as an input for the machine learning models.

Amino acid descriptors. Each row represents a different amino acid with a collection of properties, such as charge or hydrophobicity.

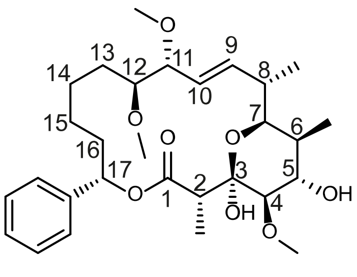
Estimating generalizability allows for the selection of the optimal protein descriptor and machine learning model.

By partitioning the data into subsets (blue and gray boxes) and measuring performance over all of them helps in estimating model performance and helps in the selection of optimal protein descriptors and machine learning models.

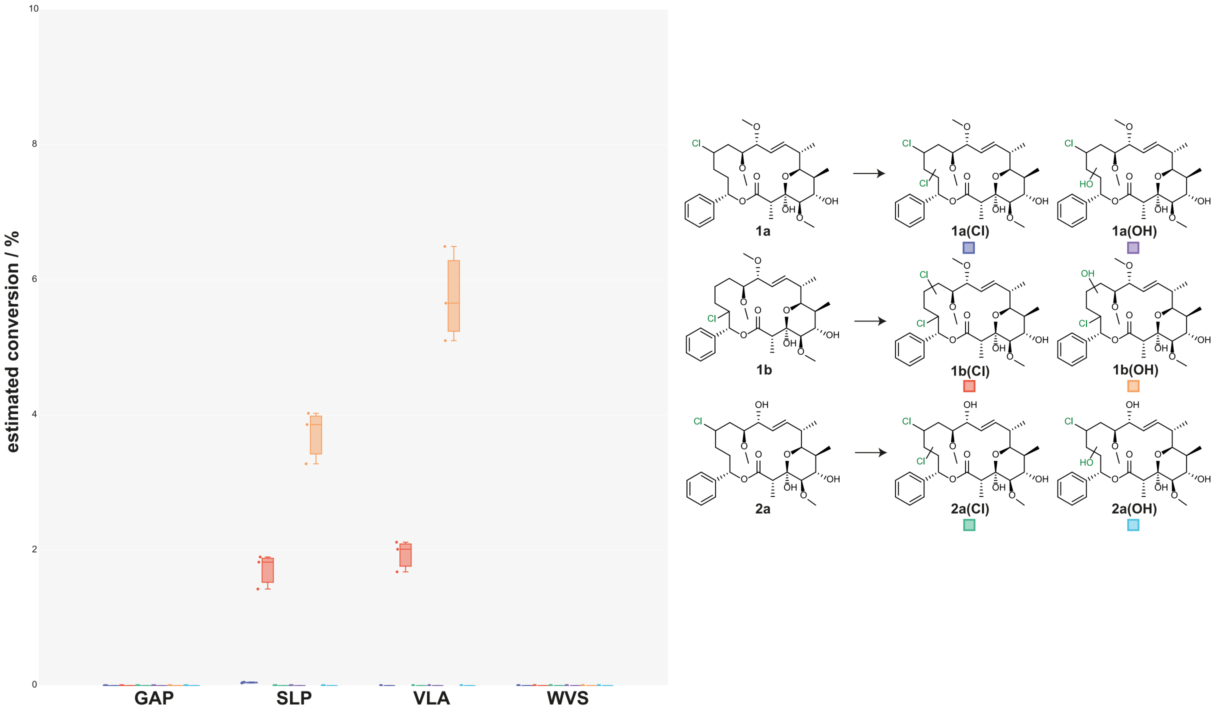
Supplementary Figure 5. Illustration of the algorithm aided approach to predict improved variants. Activity (or selectivity) data obtained by LC-MS analysis was used as a label for the machine learning algorithm. Amino acids properties were represented as a 17-dimensional vector. The feature vector of a sequence was defined by joining the vector representation of its individual amino acids at sites V81X, A88X, I161X and aggregated into the *504 x 51*-dimensional training matrix. This was used to train a machine learning model. To avoid overfitting and to better gauge the generalizability of our model, we cross-validated over ten splits, and model performance was evaluated on the coefficient of determination (R^2).



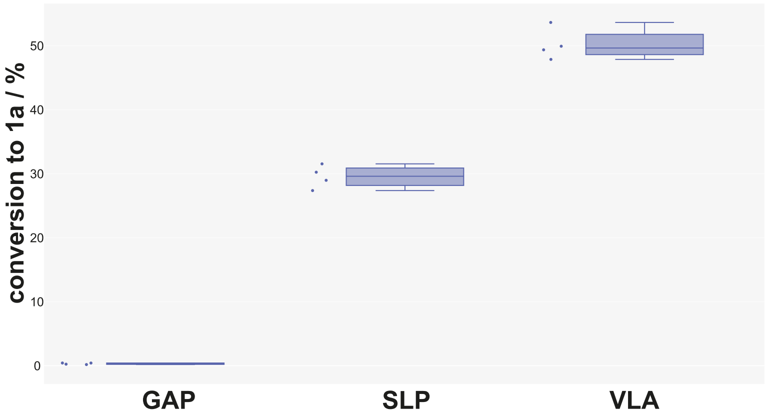
Supplementary Figure 6. Docking studies of soraphen A into WelO5* homology models. Enzyme models were prepared with SWISS-MODEL[27] (enzyme model and soraphen A in wheat, engineered residues in red) or AlphaFold[28] (enzyme model and soraphen A in palegreen, engineered residues in orange) and soraphen A was docked using AutoDock Vina[29]. In the active site the histidines coordinating to the iron (orange) are shown in grey, the chlorine in green and the α-ketoglutarate in pale cyan. **a-c** View into the active site of the WelO5* variants GAP, VLA and AHG, respectively. The active site of WelO5* including the engineered residues at position 81, 88 and 161 are nearly identical in both homology models (SWISS-MODEL (wheat) and AlphaFold (palegreen)). **d-f** Soraphen A docked into the WelO5* variants GAP, VLA and AHG. Distances were measured from the iron and chloride to C14 of soraphen A (grey dotted lines) and to C16 of soraphen A (yellow dotted lines). The two models of **d** WelO5* GAP and **f** WelO5* AHG show a similar positioning of the engineered amino acid residues as well as of the docked soraphen A. In the AlphaFold model of **e** WelO5* VLA shorter distances between the iron and chloride to C14 of soraphen A (grey dotted lines) than to C16 of the macrolide (yellow dotted lines) suggest the structural reason for the predominant formation of regioisomer **1a**. **g** Overall structural homology of the enzyme models prepared with SWISS-MODEL (wheat) or AlphaFold (palegreen): Main structural differences lay in the two α-helices marked in the black circle.



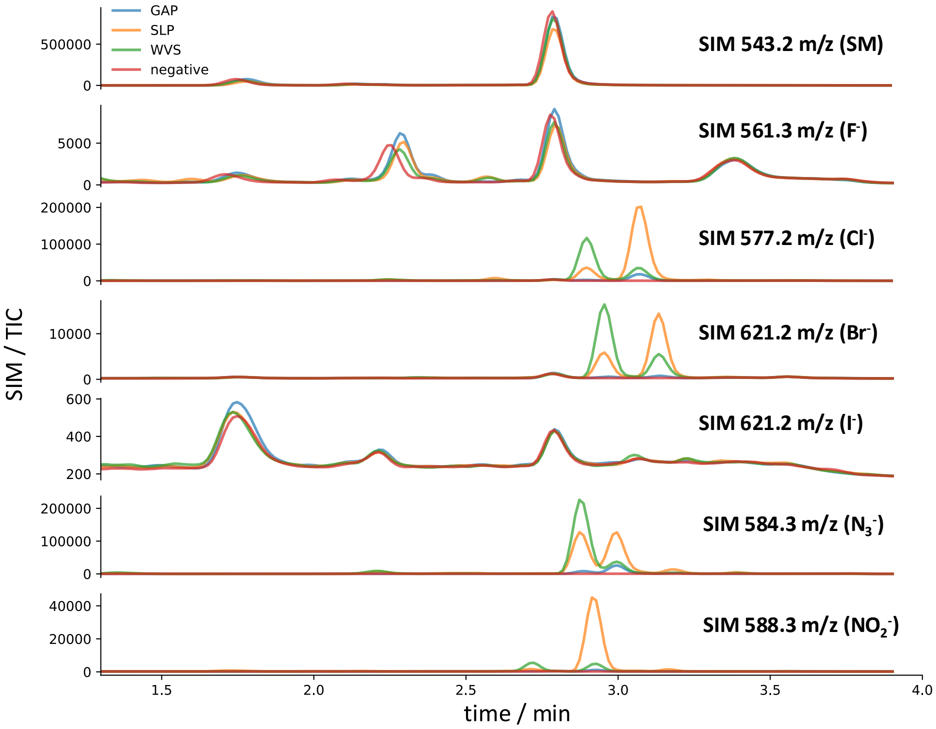
Supplementary Figure 7. Structure of soraphen A. Numbering of the carbon atoms of the cyclic polyketide backbone of soraphen A.



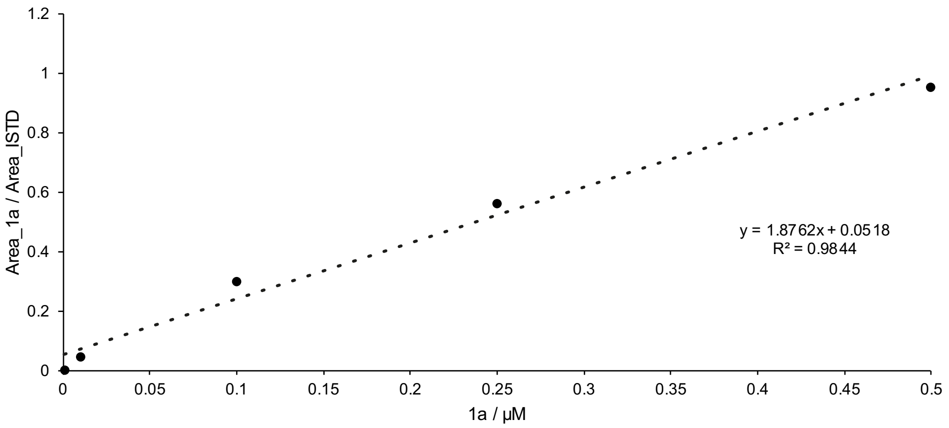
Supplementary Figure 8. *In vitro* activity assays using mono-chlorinated products (1a, 1b and 2a) as substrates for the engineered WelO5* variants GAP, SLP, VLA and WVS. Depicted is the estimated conversion to chlorinated or hydroxylated product (*SIM area of product / SIM area of all products and starting material* * 100). Reactions were performed according to the method described in the main paper using an enzyme concentration of 5 µM for variants GAP, SLP, VLA and WVS and a substrate concentration of 60 µM (1a, 1b and 2a). The conversion values to the chlorinated or hydroxylated products by the engineered halogenase variants were determined in triplicates (N = 3 independent experiments). The depicted boxes correspond to the interquartile range and end at the quartiles Q_1 and Q_3, respectively. The statistical median is depicted as a horizontal line in the box. The whiskers comprise the farthest points that are not outliers (i.e., that are within 1.5x of the interquartile range of Q_1 and Q_3, respectively).



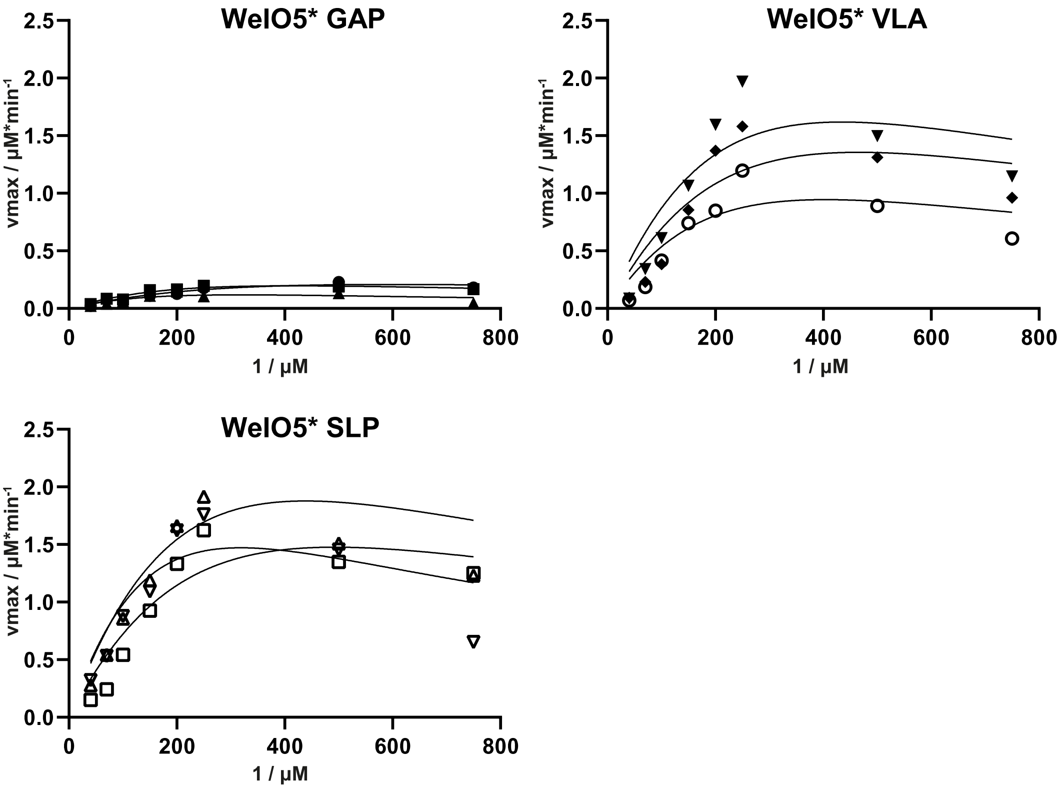
Supplementary Figure 9. *In vitro* activity assays showing the conversion of soraphen A to the chlorinated product 1a by the engineered WelO5* variants GAP, SLP and VLA. The conversion was determined at a substrate concentration of 60 µM and the reaction was quenched at stable product concentration using the procedure described in the method section of the main paper. To account for the different enzyme concentrations used (GAP = 5 µM, SLP = 0.5 µM and VLA = 0.5 µM), the observed product concentration in the GAP reactions was divided by ten. The conversion values of soraphen A to the chlorinated product **1a** by the engineered halogenase variants were determined in quadruplicates in each case (N = 4 independent experiments). The depicted boxes correspond to the interquartile range and end at the quartiles Q_1 and Q_3. The statistical median is depicted as a horizontal line in the box. The whiskers comprise the farthest points that are not outliers (i.e., that are within 1.5x the interquartile range of Q_1 and Q_3, respectively).



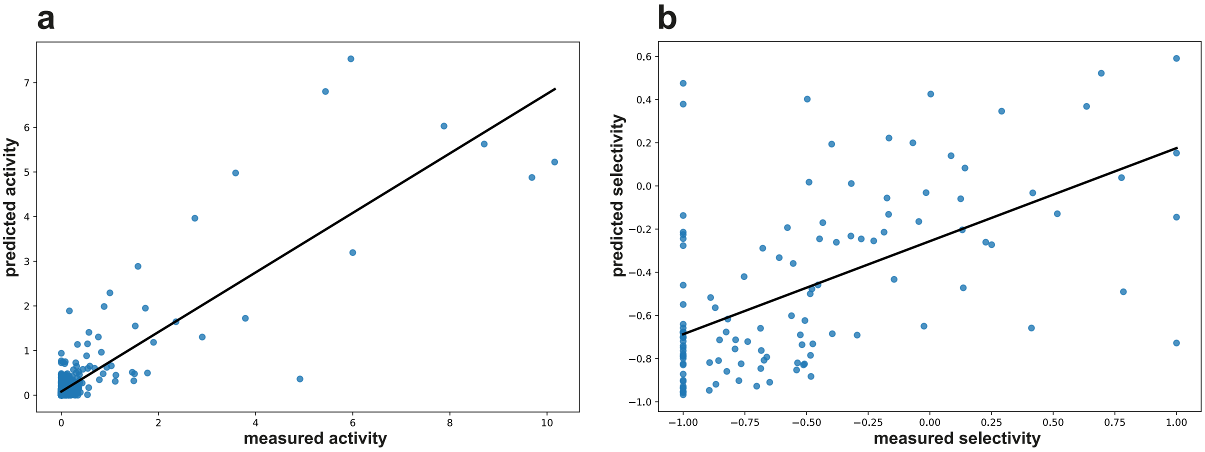
Supplementary Figure 10. LC-MS analysis of the anion promiscuity of selected WelO5* variants. The variants GAP (blue), SLP (orange), WVS (green) and negative control (no enzyme, red) were analysed in the presence of 500 mM of NaF, NaCl, NaBr, NaI, NaN_3 and $NaNO_2$, respectively, by selected ion monitoring. Masses corresponding to the introduction of the anions into soraphen A could be observed for chloride, bromide, azide and nitrite salts.



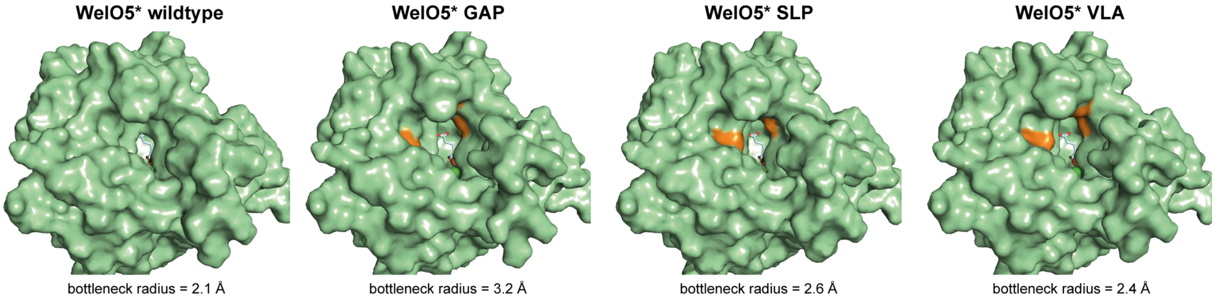
Supplementary Figure 11. Calibration curve of 1a used for product quantification.



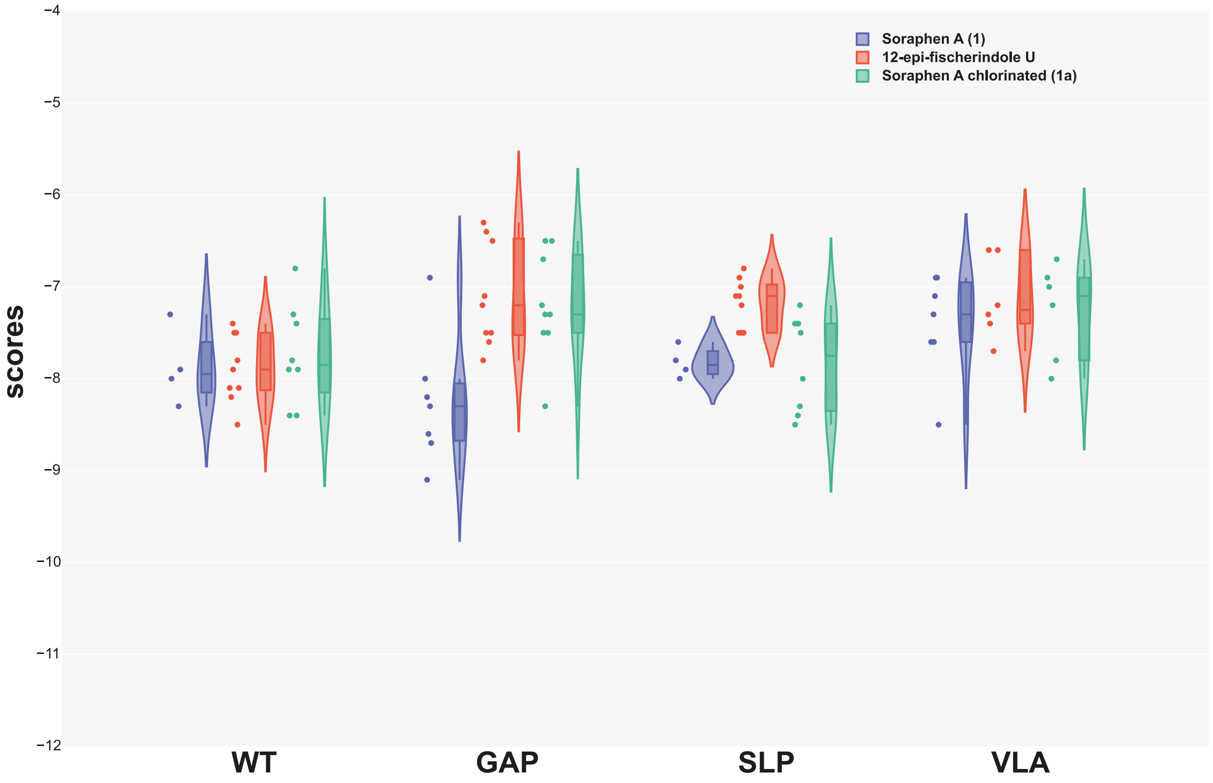
Supplementary Figure 12. Michaelis Menten kinetics for WelO5* variants GAP, VLA and SLP. The formation of product **1a** for the WelO5* variants GAP, VLA and SLP was measured in triplicate at each substrate concentration (N=3 independent experiments). All data points belonging to individual Michaelis Menten measurement series are marked by triangles, squares or circles. Substrate inhibition can be observed in all cases.



Supplementary Figure 13. Out of fold predicted vs measured values. a The measured activity values of the training set were predicted using Gaussian processes (y-axis) and compared to the measured activity (x-axis) **b** The measured selectivity of the training set were predicted using a random forest algorithm (y-axis) and compared to the measured selectivity (x-axis). A linear regression is shown for these values.


WelO5* wildtype
WelO5* GAP
WelO5* SLP
WelO5* VLA
bottleneck radius = 2.1 Å
bottleneck radius = 3.2 Å
bottleneck radius = 2.6 Å
bottleneck radius = 2.4 Å

Supplementary Figure 14. Overview of the AlphaFold models (green) of WelO5* and the variants GAP, SLP and VLA showing a view of the entrance to the active site. The engineered amino acid residues are shown in orange, while the co-factor α-ketoglutarate is shown in light cyan. To comparatively evaluate substrate access to the active sites of all enzyme variants, the bottleneck radii were calculated using CAVER Web 1.0 with default parameters[30]. This investigation highlighted that the employed enzyme engineering approach has led to a widening of the access tunnel from 2.1 Å (wildtype enzyme) to 3.2 Å (variant GAP), 2.6 Å (variant SLP) and 2.4 Å (variant VLA). The resulting improved access to the active site might explain why the wildtype enzyme cannot convert the macrolide soraphen A, while variants GAP, SLP and VLA accept the bulky substrate.



Supplementary Figure 15. Docking experiments of soraphen A, 12-epi-fischerindole U and chlorinated soraphen A into wildtype WelO5* and its engineered variants. Overview of the docking scores of soraphen A (blue, **1**), the natural substrate 12-epi-fischerindole U (red) and chlorinated soraphen A (green, **1a**) into AlphaFold models of WelO5* wild type (WT) and the variants GAP, SLP and VLA*,* respectively. The docking was performed as described in the method section of the main paper. All ligands, irrespective of them being native or non-native substrates, showed similar docking scores for the enzyme variants. Scores were obtained through the AutoDock Vina scoring function[32] which consists of the weighted sum of steric interactions (gauss$_1$, gauss$_2$ and repulsion, identical for all atom pairs), hydrophobic interaction between hydrophobic atoms and hydrogen bonding (where applicable). A lower score indicates higher affinity of the ligand towards the receptor. In each case, nine docking solutions were obtained from one docking experiment (n = 1 individual experiment). Results were visually inspected using PyMOL software and only solutions in which the ligand docked close to the active site were considered in the depicted analysis. The depicted boxes correspond to the interquartile range and end at the quartiles Q$_1$ and Q$_3$. The median is depicted as a horizontal line in the box. The whiskers comprise the farthest points that are not outliers (i.e., that are within 1.5x the interquartile range of Q$_1$ and Q$_3$, respectively).

III. Supplementary Methods

Chemical synthesis methods

Synthesis: Soraphen A was obtained by fermentation at Syngenta (former Novartis) using the published procedure[31]. The reagents for synthesis were obtained from commercial sources and used without further purification unless otherwise stated. The solvents for synthesis were obtained from commercial sources and stored over molecular sieves.

Purification: purification over silica gel were performed on a Combi*Flash* Rf 200i instrument using standard commercial pre-packed silica gel cartridges.

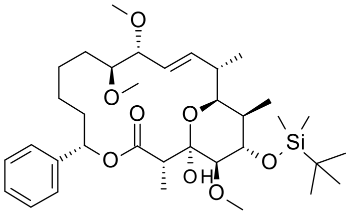
NMR: NMRs were recorded either on a Bruker 400 MHz spectrometer or a Bruker 600 MHz spectrometer. ^{1}H-NMR chemical shifts are reported relative to TMS and are referenced based on the residual proton resonances of the corresponding deuterated solvent ($CDCl_3$: 7.26 ppm) whereas ^{13}C NMR spectra are reported relative to TMS using the carbon signals of the deuterated solvent ($CDCl_3$: 77.16 ppm). Assignments were made on the basis of chemical shifts, coupling constants, COSY, HSQC, HMBC, ROESY data. Resonances are described using the following abbreviations; s (singlet), d (doublet), t (triplet), q (quartet), quin. (quintet), sext. (sextet), sept. (septet), m (multiplet), br. (broad), app. (apparent), dd (double doublet) and so on. Coupling constants (*J*) are given in Hz and are rounded to the nearest 0.1 Hz.

HPLC-MS: HPLC traces were obtained on an Acquity UPLC from Waters: Binary pump, heated column compartment, diode-array detector and ELSD detector. Column: Waters UPLC HSS T3, 1.8 µm, 30 x 2.1 mm, Temp: 60 °C, DAD Wavelength range: 210 to 500 nm, Solvent Gradient: A = water + 5% MeOH + 0.05 % HCOOH, B= Acetonitrile + 0.05 % HCOOH, gradient: 10-100% B in 2.7 min; Flow: 0.85 mL/min. Low resolution mass spectra were recorded on a mass spectrometer from Waters (SQD, SQDII Single quadrupole mass spectrometer) equipped with an electrospray source (Polarity: positive and negative ions); Capillary: 3.00 kV, Cone range: 30V, Extractor: 2.00 V, Source Temperature: 150 °C, Desolvation Temperature: 350 °C, Cone Gas Flow: 50 L/h, Desolvation Gas Flow: 650 L/h, Mass range: 100 to 900 Da.

Synthesis and characterization of soraphen C and soraphen analogues

For a scheme of the synthesis route refer to Supplementary Figure 2.

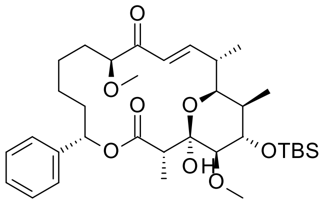
(1*R*,2*S*,5*S*,10*S*,11*R*,12*E*,14*S*,15*S*,16*R*,17*S*,18*R*)-17-[tert-butyl(dimethyl)silyl]oxy-1-hydroxy-10,11,18-trimethoxy-2,14,16-trimethyl-5-phenyl-4,19-dioxabicyclo[13.3.1]nonadec-12-en-3-one (**S1**)



Soraphen A (100 mg, 0.19 mmol), DMAP (25 mg, 0.20 mmol), imidazole (52 mg, 0.77 mmol) and TBSCl (58 mg, 0.38 mmol) were dissolved in DMF (2.0 mL). The slightly yellow solution was stirred at room temperature for 48 h. Then the reaction mixture was diluted with EtOAc and washed with aq HCl (1M), water and brine, dried over Na_2SO_4, filtered and concentrated under reduced pressure.
The crude was purified by flash chromatography on silica gel (gradient: cHex/EtOAc 95:5 to 55:45, 30 mL/min, 15 min) to afford the title compound (60 mg, 49%) as well as some recovered starting material (50 mg).

^{1}H-NMR (400 MHz, $CDCl_3$): δ (ppm) = 7.42–7.23 (m, 5H), 6.40 (dd, *J* = 16.2, 3.8 Hz, 1H), 6.12 (dd, *J* = 12.0, 2.5 Hz, 1H), 5.41 (ddd, *J*=16.0, 9.5, 1.8 Hz, 1H), 5.17 (d, *J*=1.8 Hz, 1H), 4.17 (t, *J*=2.5 Hz, 1H), 3.78 (dd, *J*=9.5, 2.2 Hz, 1H), 3.69 (dd, *J*=10.5, 2.5 Hz, 1H), 3.47 (s, 3H), 3.42 (dt, *J*=11.1, 2.5 Hz, 1H), 3.38 (s, 3H), 3.31 (s, 3H), 3.05 (qd, *J*=7.0, 1.1 Hz, 1H), 3.00 (dd, *J*=2.7, 0.9 Hz, 1H), 2.56–2.49 (m, 1H), 2.13–2.04 (m, 1H), 1.81–1.72 (m, 2H), 1.65–1.54 (m, 3H), 1.34–1.25 (m, 2H), 1.16 (d, *J*=7.3 Hz, 3H), 1.09 (d, *J*=7.3 Hz, 3H), 1.05–1.01 (m, 1H), 1.00 (d, *J*=6.5 Hz, 3H), 0.94 (s, 9H), 0.17 (d, *J*=4.0 Hz, 6H);
^{13}C-NMR (101 MHz, $CDCl_3$): δ (ppm) = 171.25, 142.77, 140.74, 128.43 (2C), 127.52, 126.29 (2C), 121.80, 99.45, 84.99, 83.65, 77.39, 72.08, 71.90, 70.86, 58.32, 57.41, 56.31, 46.32, 37.30, 35.62, 35.22, 31.02, 25.95 (3C), 25.23, 24.18, 18.26, 12.42, 11.49, 10.32, –4.82, –4.85.

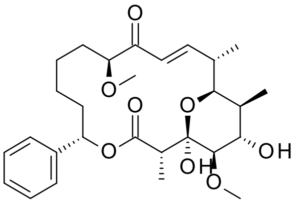
(1*R*,2*S*,5*S*,10*S*,12*E*,14*S*,15*S*,16*R*,17*S*,18*R*)-17-[tert-butyl(dimethyl)silyl]oxy-1-hydroxy-10,18-dimethoxy-2,14,16-trimethyl-5-phenyl-4,19-dioxabicyclo[13.3.1]nonadec-12-ene-3,11-dione (**S2**)



To a solution of **S1** (750 mg, 1.18 mmol) in 1,2-dichloroethane (14.8 mL) at room temperature was added DDQ (621 mg, 2.60 mmol). The yellow suspension was stirred for 4 days. Then the reaction mixture was diluted with EtOAc, washed twice with aq. $Na_2S_2O_3$ (10%), then brine, dried over Na_2SO_4, filtered and concentrated under reduced pressure.
The crude was purified by flash chromatography on silica gel (gradient: cHex/EtOAc 90:10 to 80:20, 40 mL/min, 16 min) to afford the title compound (522.8 mg, 72%).

^{1}H-NMR (400 MHz, $CDCl_3$): δ (ppm) = 7.45 (dd, *J*=16.9, 4.8 Hz, 1H), 7.39–7.28 (m, 4H), 6.24 (dd, *J*=16.9, 1.5 Hz, 1H), 5.91 (dd, *J*=8.6, 5.3 Hz, 1H), 5.05 (d, *J*=1.5 Hz, 1H), 4.29 (dd, *J*=7.7, 5.5 Hz, 1H), 4.16 (t, *J*=2.6 Hz, 1H), 3.86 (dd, *J*=10.3, 2.6 Hz, 1H), 3.37 (s, 6H), 3.05 (qd, *J*=7.0, 1.47 Hz, 1H), 3.00 (dd, *J*=2.8, 0.9 Hz, 1H), 2.70–2.64 (m, 1H), 2.03–1.94 (m, 1H), 1.82–1.71 (m, 2H), 1.70–1.58 (m, 4H), 1.49–1.39 (m, 3H), 1.13 (d, *J*=7.0 Hz, 3H), 1.10 (d, *J*=7.3 Hz, 3H), 1.00 (d, *J*=6.6 Hz, 3H), 0.94 (s, 9H), 0.16 (d, *J*=5.1 Hz, 6H);
^{13}C-NMR (101 MHz, $CDCl_3$): δ (ppm) = 202.74, 171.44, 152.37, 141.78, 128.45 (2C), 127.67, 127.22, 126.64 (2C), 99.90, 83.63, 74.30, 70.76, 70.61, 57.92, 57.50, 46.31, 36.52, 36.01, 35.73, 32.33, 25.92 (3C), 23.98, 23.49, 18.23, 13.20, 11.55, 10.24, –4.82, –4.86;
HPLC-MS: rt = 2.69 min, m/z = 504 $[M-C_6H_{14}Si]^-$; 618 $[M-H]^-$

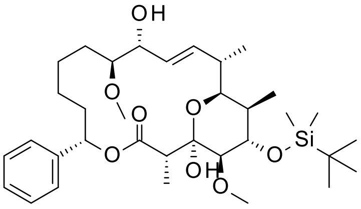
2*S*,5*S*,10*S*,12*E*,14*S*,15*S*,16*S*,17*S*,18*R*)-1,17-dihydroxy-10,18-dimethoxy-2,14,16-trimethyl-5-phenyl-4,19-dioxabicyclo[13.3.1]nonadec-12-ene-3,11-dione (**3**)



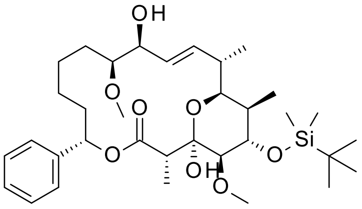
To a solution of **S2** (30.0 mg, 0.485 mmol) in THF (0.2 mL) were added at 0°C AcOH (8.8 µL, 0.15 mmol) and TBAF (0.10 mL, 1 M in THF, 0.10 mmol). The reaction mixture was stirred at 0 °C for 1.5 h. The reaction mixture was poured into sat. aq. NH_4Cl, then extracted with DCM/MeOH (9:1) The combined organic layers were dried over Na_2SO_4, filtered and concentrated under reduced pressure.
The crude product was purified by flash chromatography on silica gel (gradient: cHex/EtOAc 85:15 to 50:50, 18 mL/min, 12 min) to afford the title compound (21 mg, 86%).

^{1}H-NMR (400 MHz, $CDCl_3$): δ (ppm) = 7.35–7.30 (m, 6H), 6.51 (dd, *J*=16.1, 1.5 Hz, 1H), 5.61 (t, *J*=7.3 Hz, 1H), 4.51 (s, 1H), 4.04–4.01 (m, 2H), 3.80 (dd, *J*=9.2, 4.4 Hz, 1H), 3.38 (s, 3H), 3.35 (s, 3H), 3.17–3.12 (m, 2H), 2.69–2.60 (m, 1H), 1.99–1.93 (m, 1H), 1.90–1.84 (m, 2H), 1.80–1.60 (m, 3H), 1.51–1.41 (m, 2H), 1.36–1.29 (m, 1H), 1.25–1.17 (m, 1H), 1.07 (d, *J*=7.3 Hz, 3H), 1.06 (d, *J*=7.3 Hz, 3H), 1.04 (d, *J*=7.0 Hz, 3H);
^{13}C-NMR (101 MHz, $CDCl_3$): δ (ppm) = 202.60, 170.99, 152.95, 140.15, 128.67 (2C), 128.31, 126.68 (2C), 124.57, 99.85, 86.02, 76.62, 76.16, 72.63, 68.99, 57.76, 57.62, 45.76, 36.53, 35.65, 34.44, 30.18, 23.74, 22.77, 13.94, 11.79, 10.49;
HPLC-MS: rt = 1.83 min, m/z = 504 $[M-H]^-$; 528 $[M+H+Na]^{2+}$

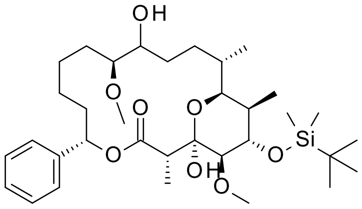
(1*R*,2*S*,5*S*,10*S*,11*R*,12*E*,14*S*,15*S*,16*R*,17*S*,18*R*)-17-[tert-butyl(dimethyl)silyl]oxy-1,11-dihydroxy-10,18-dimethoxy-2,14,16-trimethyl-5-phenyl-4,19-dioxabicyclo[13.3.1]nonadec-12-en-3-one (S3)
(1*R*,2*S*,5*S*,10*S*,11*S*,12*E*,14*S*,15*S*,16*R*,17*S*,18*R*)-17-[tert-butyl(dimethyl)silyl]oxy-1,11-dihydroxy-10,18-dimethoxy-2,14,16-trimethyl-5-phenyl-4,19-dioxabicyclo[13.3.1]nonadec-12-en-3-one (S4)
(1*R*,2*S*,5*S*,10*S*,14*S*,15*S*,16*R*,17*S*,18*R*)-17-[tert-butyl(dimethyl)silyl]oxy-1,11-dihydroxy-10,18-dimethoxy-2,14,16-trimethyl-5-phenyl-4,19-dioxabicyclo[13.3.1]nonadecan-3-one (**S5**)



S3



S4



S5

To a solution of **S2** (380 mg, 0.610 mmol) in 1,2-dimethoxyethane (3 mL) was added at 0°C a solution of $ZnCl_2$ (1 M in Et_2O, 9.0 mL, 0.92 mmol) followed by addition of $NaBH_4$ (64.5 mg, 1.53 mmol). The reaction mixture was stirred at 0 °C for 1h40. It was then quenched with sat. aq. NH_4Cl and extracted with EtOAc. The combined organic phases were dried over Na_2SO_4, filtered and concentrated under reduced pressure.
The crude product was purified by flash chromatography on silica gel (gradient: cHex/EtOAc 95:05 to 70:30, 40 mL/min, 22 min) to afford three products: **S3** (186.1 mg, 49%), **S4** (78.4 mg, 21%), and **S5** (23.9 mg, 6%) as a mixture of epimers.

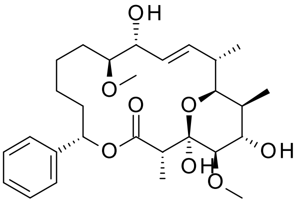
Data for S3:
^{1}H-NMR (400 MHz, $CDCl_3$): δ (ppm) = 7.37–7.22 (m, 5H), 6.33 (dd, *J*=16.1, 4.0 Hz, 1H), 6.07 (dd, *J*=12.1, 2.6 Hz, 1H), 5.39 (ddd, *J*=16.1, 9.4, 1.7 Hz, 1H), 5.17 (d, *J*=1.8 Hz, 1H), 4.24 (dd, *J*=9.2, 2.6 Hz, 1H), 4.14 (t, *J*=2.8 Hz, 1H), 3.64 (dd, *J*=10.6, 2.6 Hz, 1H), 3.44 (s, 3H), 3.36 (s, 3H), 3.34–3.33 (m, 1H), 3.03 (qd, *J*=7.0, 1.5 Hz, 1H), 2.98 (dd, *J*=2.9, 1.1 Hz, 1H), 2.52–2.43 (m, 1H), 2.29 (br m, 1H), 2.13–2.07 (m, 1H), 1.79–1.56 (m, 5H), 1.32–1.24 (m, 2H), 1.21–1.17 (m, 1H), 1.13 (d, *J*=7.3 Hz, 3H), 1.06 (d, *J*=7.7 Hz, 3H), 0.95 (d, *J*=6.6 Hz, 3H), 0.93 (s, 9H), 0.15 (s, 3H), 0.14 (s, 3H);
HPLC-MS: rt = 2.58 min, m/z = 644 $[M+H+Na]^{2+}$; 506 $[M-C_6H_{14}Si]^-$

The relative stereochemistry of **S3** was determined retrospectively from soraphen C after deprotection. The relative stereochemistry of **S4** was determined by comparison with **S3**.

Data for S4:
^{1}H-NMR (400 MHz, $CDCl_3$): δ (ppm) = 7.38–7.24 (m, 5H), 6.12 (ddd, *J*=16.3, 6.1, 1.1 Hz, 1H), 5.90 (dd, *J*=8.6, 5.7 Hz, 1H), 5.55 (ddd, *J*=16.2, 5.8, 1.1 Hz, 1H), 5.08 (d, *J*=1.1 Hz, 1H), 4.29 (t, *J*=5.7 Hz, 1H), 4.11 (t, *J*=2.6 Hz, 1H), 3.84 (dd, *J*=9.9, 2.6 Hz, 1H), 3.43 (s, 3H), 3.35 (s, 3H), 3.35–3.30 (m, 1H), 3.03 (qd, *J*=7.1, 1.1 Hz, 1H), 2.97 (dd, *J*=2.9, 0.7 Hz, 1H), 2.70–2.58 (br s, 1H), 2.53–2.46 (m, 1H), 2.09–2.00 (m, 1H), 1.75–1.62 (m, 5H), 1.56–1.47 (m, 2H), 1.38–1.29 (m, 2H), 1.11 (d, *J*=7.0 Hz, 3H), 1.07 (d, *J*=7.3 Hz, 3H), 0.97 (d, *J*=7.0 Hz, 3H), 0.93 (m, 9H), 0.15 (s, 3H), 0.14 (s, 3H);
^{13}C-NMR (101 MHz, $CDCl_3$): δ (ppm) = 171.91, 141.78, 136.68, 128.43 (2C), 127.65, 126.96, 126.71 (2C), 99.54, 83.58, 77.68, 74.18, 73.18, 71.62, 70.95, 57.69, 57.52, 46.66, 36.69, 36.22, 35.79, 28.91, 25.92 (3C), 24.47, 23.97, 18.23, 15.63, 11.72, 10.74, –4.83, –4.85;
HPLC-MS: rt = 2.66 min, m/z = 644 [M+H+Na]$^{2+}$

Data for S5:
^{1}H-NMR (400 MHz, $CDCl_3$): δ (ppm) = 7.39–7.28 (m, 4H), 7.25–7.21 (m, 1H), 5.98 (dd, *J*=10.8, 3.1 Hz, 1H), 5.15 (d, *J*=1.8 Hz, 1H), 4.15 (t, *J*=2.6 Hz, 1H), 3.88 (dd, *J*=10.6, 2.6 Hz, 1H), 3.74 (dt, *J*=9.1, 3.0 Hz, 1H), 3.42 (s, 3H), 3.39–3.36 (m, 1H), 3.34 (s, 3H), 3.01–2.94 (m, 2H), 2.12–1.98 (m, 3H), 1.93–1.75 (m, 2H), 1.68– 1.56 (m, 6H), 1.53–1.38 (m, 4H), 1.12 (d, *J*=7.0 Hz, 3H), 1.03 (d, *J*=7.3 Hz, 3H), 0.94 (s, 9H), 0.81 (d, *J*=7.0 Hz, 3H), 0.16 (s, 3H), 0.15 (s, 3H);
^{13}C-NMR (101 MHz, $CDCl_3$): δ (ppm) = 171.14, 142.53, 128.43 (2C), 127.58, 126.49 (2C), 99.61, 83.92, 72.81, 72.24, 70.93, 67.36, 57.76, 57.28, 46.52, 36.52, 35.11, 32.06, 28.98, 27.79, 25.94 (3C), 25.15, 23.09, 22.92, 18.25, 14.67, 11.56, 10.45, –4.82 (2C);
HPLC-MS: rt = 2.63 min, m/z = 646 [M+H+Na]$^{2+}$; 508 [M-$C_6H_{14}Si$]$^-$

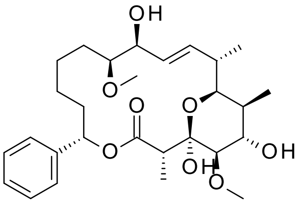
(1*R*,2*S*,5*S*,10*S*,11*R*,12*E*,14*S*,15*S*,16*S*,17*S*,18*R*)-1,11,17-trihydroxy-10,18-dimethoxy-2,14,16-trimethyl-5-phenyl-4,19-dioxabicyclo[13.3.1]nonadec-12-en-3-one (soraphen C, **2**)



To a solution of **S3** (271 mg, 0.535 mmol) in THF (4 mL) were added at 0°C AcOH (75.8 µL, 1.31 mmol) and TBAF (0.87 mL, 1 M in THF, 0.87 mmol). The reaction mixture was stirred at 0°C for 2 h then at room temperature for 18 h. The reaction mixture was poured into sat. aq. NH_4Cl, then extracted with EtOAc. The combined organic layers were dried over Na_2SO_4, filtered and concentrated under reduced pressure.
The crude product was purified by flash chromatography on silica gel (gradient: cHex/EtOAc 90:10 to 40:60, 35 mL/min, 18 min) to afford the title compound (219 mg, 99%).
Comparison of the spectral data with the ones of the isolated natural product[32,33] confirmed the identity of the product. A table of compared ^{13}C-NMR shifts is presented in Supplementary Table 7.

^{1}H-NMR (400 MHz, $CDCl_3$): δ (ppm) = 7.37–7.28 (m, 5H), 6.15 (dd, *J*=16.0, 3.9 Hz, 1H), 5.82 (dd, *J*=11.2, 3.5 Hz, 1H), 5.48 (ddd, *J*=16.0, 9.4, 1.8 Hz, 1H), 4.36 (s, 1H), 4.17 (td, *J*=9.0, 2.6 Hz, 1H), 4.01 (br d, *J*=8.0 Hz, 1H), 3.81 (dd, *J*=10.4, 2.8 Hz, 1H), 3.62 (d, *J*=9.9 Hz, 1H), 3.43 (s, 3H), 3.38 (s, 3H), 3.37–3.34 (m, 1H), 3.18 (dd, *J*=2.6, 1.1 Hz, 1H), 3.14 (q, *J*=7.3 Hz, 1H), 2.50–2.42 (m, 2H), 2.15–2.07 (m, 1H), 1.93 (q, *J*=7.0 Hz, 1H), 1.82–1.75 (m, 1H), 1.72–1.64 (m, 1H), 1.51–1.45 (m, 2H), 1.38–1.34 (m, 1H), 1.22–1.13 (m, 2H), 1.09 (d, *J*=7.0 Hz, 3H), 1.05 (d, *J*=7.7 Hz, 3H), 1.00 (d, *J*=6.6 Hz, 3H);
^{13}C-NMR (101 MHz, $CDCl_3$): δ (ppm) = 170.78, 141.12, 137.45, 128.72 (2C), 128.29, 126.35 (2C), 125.18, 99.60, 83.89, 76.28, 75.04, 74.81, 72.66, 68.98, 57.78, 57.47, 46.35, 36.01, 35.79, 35.32, 29.57, 26.14, 23.17, 12.64, 11.83, 10.49;
HPLC-MS: rt = 1.65 min, m/z = 505 [M-H]$^-$
HR-MS: m/z calculated for $C_{28}H_{43}O_8$ [(M+H)$^+$]: 507.2952, found 507.2962.

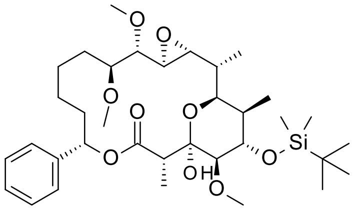
(1*R*,2*S*,5*S*,10*S*,11*S*,12*E*,14*S*,15*S*,16*S*,17*S*,18*R*)-1,11,17-trihydroxy-10,18-dimethoxy-2,14,16-trimethyl-5-phenyl-4,19-dioxabicyclo[13.3.1]nonadec-12-en-3-one (*epi*-soraphen C, **S7**)



To a solution of **S4** (10 mg, 0.016 mmol) in THF (0.4 mL) were added at 0°C AcOH (2.8 µL, 0.048 mmol) and TBAF (0.03 mL, 1 M in THF, 0.32 mmol). The reaction mixture was stirred at 0°C for 2 h then at room temperature for 18 h. The reaction mixture was poured into sat. aq. NH_4Cl, then extracted with a mixture of DCM/MeOH (9:1). The combined organic layers were dried over Na_2SO_4, filtered and concentrated under reduced pressure.
The crude product was purified by flash chromatography on silica gel (gradient: cHex/EtOAc 70:30 to 50:50, 18 mL/min, 9 min) to afford the title compound (5 mg, 61%) containing some impurities.

^{1}H-NMR (600 MHz, $CDCl_3$): δ (ppm) = 7.37–7.29 (m, 5H), 5.89 (dd, *J*=15.9, 6.3 Hz, 1H), 5.67 (t, *J*=7.3 Hz, 1H), 5.53 (ddd, *J*=15.8, 7.1, 1.1 Hz, 1H), 4.69 (s, 1H), 4.20 (t, *J*=6.6 Hz, 1H), 3.99 (dd, *J*=10.4, 2.5 Hz, 1H), 3.98 (br. s, 1H), 3.69 (br. s, 1H), 3.42 (s, 3H), 3.37 (s, 3H), 3.30–3.26 (m, 1H), 3.14 (dd, *J*=2.5, 0.9 Hz, 1H), 3.11 (q, *J*=7.2 Hz, 1H), 2.53 (br s, 1H), 2.47–2.43 (m, 1H), 2.24–2.18 (m, 1H), 1.94–1.90 (m, 1H), 1.85–1.78 (m, 2H), 1.66–1.61 (m, 1H), 1.56–1.51 (m, 1H), 1.46–1.38 (m, 3H), 1.04 (d, *J*=7.5 Hz, 3H), 1.03 (d, *J*=7.1 Hz, 3H), 0.99 (d, *J*=6.6 Hz, 3H);
^{13}C-NMR (151 MHz, $CDCl_3$): δ (ppm) = 172.82, 139.69, 137.96, 128.76 (2C), 128.70, 128.46 (2C), 127.09, 99.91, 82.86, 77.57, 76.33, 73.58, 71.74, 68.93, 57.49, 57.45, 45.42, 36.84, 35.46, 35.02, 27.58, 25.12, 22.57, 16.57, 12.55, 10.55;
HPLC-MS: rt = 1.77 min, m/z = 505 $[M-H]^-$

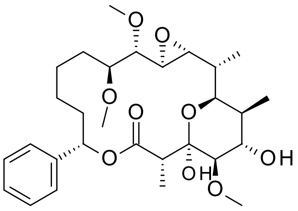
(1*S*,2*R*,3*R*,5*S*,6*S*,7*S*,12*S*,15*S*,16*R*,17*R*,18*S*,19*R*)-18-[tert-butyl(dimethyl)silyl]oxy-16-hydroxy-6,7,17-trimethoxy-2,15,19-trimethyl-12-phenyl-4,13,20-trioxatricyclo[14.3.1.03,5]icosan-14-one (**S6**)



To a solution of **S3** (400 mg, 0.630 mmol) in DCM (6.3 mL) at rt was added mCPBA (706 mg, 3.150 mmol). The reaction mixture was stirred for 17 h. The reaction mixture was poured into sat. aq. $NaHCO_3$. The phases were separated, the aqueous layer was extracted with EtOAc. The combined organic layers were washed with brine, dried over Na_2SO_4, filtered and concentrated under reduced pressure.
The crude product was purified by flash chromatography on silica gel (gradient: cHex/EtOAc 100:0 to 80:20, 40 mL/min, 18min) to afford the title compound (330 mg, 80%).

^{1}H-NMR (400 MHz, $CDCl_3$): δ (ppm) = 7.37–7.32 (m, 4H), 7.26–7.23 (m, 1H), 6.06 (dd, *J*=12.3, 2.8 Hz, 1H), 5.43 (d, *J*=1.8 Hz, 1H), 4.18 (t, *J*=1.8 Hz, 1H), 4.00 (dd, *J*=11.0, 2.6 Hz, 1H), 3.55 (s, 3H), 3.49 (t, *J*=2.0 Hz, 1H), 3.41 (s, 3H), 3.38 (s, 3H), 3.37–3.35 (m, 1H), 3.16 (dd, *J*=8.1, 2.2 Hz, 1H), 3.06–3.01 (m, 2H), 2.95 (dd, *J*=8.3, 1.7 Hz, 1H), 2.37–2.29 (m, 1H), 2.06–1.97 (m, 2H), 1.72–1.66 (m, 1H), 1.64–1.55 (m, 2H), 1.47–1.44 (m, 1H), 1.40–1.28 (m, 2H), 1.18 (d, *J*=7.3 Hz, 3H), 1.08 (d, *J*=7.3 Hz, 3H), 1.04–0.97 (m, 1H), 0.94 (s, 9H), 0.60 (d, *J*=7.0 Hz, 3H), 0.17 (s, 3H), 0.15 (s, 3H);
^{13}C-NMR (101 MHz, $CDCl_3$): δ (ppm) = 170.83, 142.61, 128.48 (2C), 127.55, 126.20 (2C), 99.68, 83.66, 83.23, 71.87, 70.52, 68.14, 58.58, 58.07, 57.39, 55.13, 53.63, 46.39, 36.89, 35.49, 34.05, 30.39, 27.06, 25.80 (3C), 24.83, 24.18, 18.05, 11.26, 10.09, 8.92, –4.86, –4.99;
HPLC-MS: rt = 2.70 min, m/z = 674 $[M+Na]^+$
HR-MS: m/z calculated for $C_{35}H_{58}NaO_9Si$ [$(M+Na)^+$]: 673.3742, found 673.3727.

(1*R*,2*R*,3*R*,5*S*,6*S*,7*S*,12*S*,15*S*,16*R*,17*R*,18*S*,19*S*)-16,18-dihydroxy-6,7,17-trimethoxy-2,15,19-trimethyl-12-phenyl-4,13,20-trioxatricyclo[14.3.1.03,5]icosan-14-one (**4**)



To a solution of **S6** (660 mg, 1.01 mmol) in THF (10 mL) were added at 0°C AcOH (176 µL, 3.04 mmol) and TBAF (2.00 mL, 1 M in THF, 2.03 mmol). The reaction mixture was stirred at 0°C for 2.5 h. The reaction mixture was poured into sat. aq. NH_4Cl, then extracted with EtOAc. The combined organic layers were dried over Na_2SO_4, filtered and concentrated under reduced pressure.

The crude product was purified by flash chromatography on silica gel (gradient: cHex/EtOAc 85:15 to 60:40, 18 mL/min, 11 min) to afford the title compound (500 mg, 92%).

^{1}H-NMR (400 MHz, $CDCl_3$): δ (ppm) = 7.35–7.24 (m, 5H), 5.86 (dd, *J*=12.3, 1.8 Hz, 1H), 4.47 (s, 1H), 4.11–4.08 (m, 2H), 3.59 (d, *J*=8.0 Hz, 1H), 3.54 (s, 3H), 3.43 (s, 3H), 3.40–3.38 (m, 1H), 3.37 (s, 3H), 3.30 (t, *J*=2.0 Hz, 1H), 3.17 (d, *J*=1.8 Hz, 1H), 3.13–3.07 (m, 2H), 2.97 (dd, *J*=8.4, 3.6 Hz, 1H), 2.32–2.25 (m, 1H), 2.10–2.04 (m, 1H), 1.87–1.81 (m, 1H), 1.73–1.62 (m, 3H), 1.55–1.44 (m, 3H), 1.25–1.20 (m, 1H), 1.14 (d, *J*=6.9 Hz, 3H), 1.05 (d, *J*=7.6 Hz, 3H), 0.64 (d, *J*=6.9 Hz, 3H);

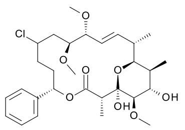
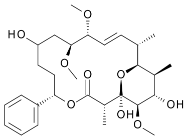
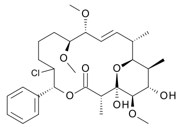
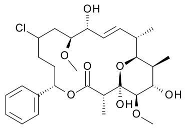
^{13}C-NMR (101 MHz, $CDCl_3$): δ (ppm) = 170.72, 141.64, 128.59 (2C), 127.99, 126.14 (2C), 99.78, 82.98, 76.34, 74.08, 68.96, 68.87, 58.75, 58.40, 57.35, 55.78, 54.43, 53.53, 46.67, 36.47, 35.40, 34.02, 29.47, 25.78, 23.73, 11.58, 10.25, 9.01;

HPLC-MS: rt = 1.86 min, m/z = 535 $[M-H]^-$

HR-MS: m/z calculated for $C_{29}H_{44}NaO_9$ $[(M+Na)^+]$: 559.2878, found 559.2882.

Purification of bio-extracts

Supplementary Table 8. Structures of isolated modified soraphen compounds.

Starting material	Structure of isolated compounds		
Soraphen A	 **1a**	 **1c**	 **1b**
Soraphen C	 **2a**		

All bioextracts were purified by preparative reverse-phase HPLC. In each case, the bioextract was dissolved in 1 mL of DMSO and injected on a Fraction Lynx Prep HPLC equipped with a Column Hichrom C18 ODS-2 5 250 mm x 2.1mm i.d. The mobile phase consisted of a mixture of H_2O+0.1% HCOOH and ACN+0.1% HCOOH. The flow was set at 20 mL min^{-1}. The gradients used for the purification of the different extracts are presented below.

Supplementary Table 9. Prep. HPLC gradient for **1a** and **1c**.

Time (min)	% H_2O + 0.1% HCOOH	% ACN + 0.1% HCOOH
0	60	40
2	60	40
37	0	100
40	1	100
41	50	50
45	50	50

(1a) rt = 23.6 min; **(1c)** rt = 11.7 min;

Supplementary Table 10. Prep. HPLC gradient for **1b**.

First column:

Time (min)	% H_2O + 0.1% HCOOH	% ACN + 0.1% HCOOH
0	60	40
2	60	40
35	0	100
40	0	100
41	60	40
45	60	40

Second column:

Time (min)	% H_2O + 0.1% HCOOH	% ACN + 0.1% HCOOH
0	50	50
2	50	50
35	0	100
40	0	100
41	50	50
45	50	50

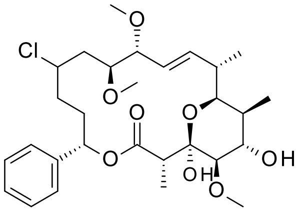
(1b) rt (second column)= 20.6 min.

Supplementary Table 11. Prep. HPLC gradient for **2a**.

Time (min)	% H_2O + 0.1% HCOOH	% ACN + 0.1% HCOOH
0	60	40
2	60	40
37	0	100
40	0	100
41	60	40
45	60	40

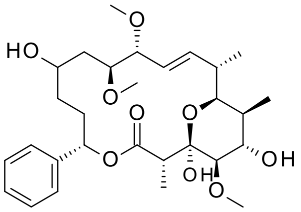
Analytical data of the halogenated and hydroxylated products

(1*R*,2*S*,5*S*,10*S*,11*R*,12*E*,14*S*,15*S*,16*S*,17*S*,18*R*)-8-chloro-1,17-dihydroxy-10,11,18-trimethoxy-2,14,16-trimethyl-5-phenyl-4,19-dioxabicyclo[13.3.1]nonadec-12-en-3-one (**1a**)



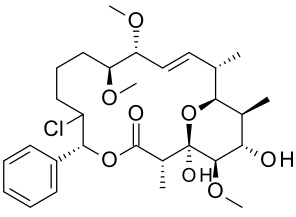
Isolated yield: 10% (11.0 mg)
^{1}H-NMR (600 MHz, CDCl$_3$) δ (ppm) = 7.36–7.34 (m, 4H), 7.30 (dq, *J*=8.5, 4.3 Hz, 1H), 6.21 (dd, *J*=16.3, 3.7 Hz, 1H), 5.92 (dd, *J*=11.8, 2.5 Hz, 1H), 5.49 (ddd, *J*=16.2, 9.5, 1.7 Hz, 1H), 4.57 (s, 1H), 4.12–4.06 (m, 1H), 4.00 (tt, *J*=11.7, 2.5 Hz, 1H), 3.88 (dt, *J*=10.7, 2.7 Hz, 1H), 3.77 (dd, *J*=10.5, 2.5 Hz, 1H), 3.68 (dd, *J*=9.5, 1.9 Hz, 1H), 3.48 (s, 3H), 3.38 (s, 3H), 3.32 (s, 3H), 3.19-3.15 (m, 2H), 3.12 (q, *J*=7.0 Hz, 1H), 2.53–2.49 (m, 1H), 2.21–2.15 (m, 1H), 2.06–2.02 (m, 1H), 2.01–1.96 (m, 1H), 1.92–1.89 (m, 1H), 1.88–1.85 (m, 1H), 1.77–1.73 (m, 1H), 1.69 (ddd, *J*=13.9, 11.5, 2.1 Hz, 1H), 1.56 (br s, 1H), 1.13 (d, *J*=7.1 Hz, 3H), 1.07 (d, *J*=7.6 Hz, 3H), 1.02 (d, *J*=6.7 Hz, 3H);
^{13}C-NMR (151 MHz, CDCl$_3$) δ = 170.97, 141.00, 140.12, 128.75 (2C), 128.30, 126.15 (2C), 122.97, 99.54, 84.03, 79.98, 76.33, 73.15, 72.40, 69.22, 58.48, 57.47, 56.89, 56.51, 46.49, 39.67, 35.85, 35.42, 34.33, 32.90, 12.58, 11.71, 10.43 ppm;
HPLC-MS: rt = 2.07 min, m/z = 578/580 $[M+H+Na]^{2+}$, 553/555 $[M–H]^-$ chloro isotopic pattern

(1*R*,2*S*,5*S*,10*S*,11*R*,12*E*,14*S*,15*S*,16*S*,17*S*,18*R*)-1,8,17-trihydroxy-10,11,18-trimethoxy-2,14,16-trimethyl-5-phenyl-4,19-dioxabicyclo[13.3.1]nonadec-12-en-3-one (**1c**)



Isolated yield: 3% (3.7 mg)
^{1}H-NMR (600 MHz, CDCl$_3$) δ = 7.36–7.33 (m, 4H), 7.30 (dq, *J*=8.7, 4.2 Hz, 1H), 6.16 (dd, *J*=16.3, 3.7 Hz, 1H), 5.87 (dd, *J*=10.2, 4.7 Hz, 1H), 5.49 (ddd, *J*=16.2, 9.3, 1.8 Hz, 1H), 4.48 (s, 1H), 4.05 (br d, *J*=8.7 Hz, 1H), 3.86 (br s, 1H), 3.81 (br dd, *J*=10.5, 2.5 Hz, 1H), 3.76–3.73 (m, 1 H), 3.72 (dd, *J*=9.3, 1.8 Hz, 1H), 3.47 (s, 3H), 3.42 (br d, *J*=9.3 Hz, 1H), 3.38 (s, 3H), 3.30 (s, 3H), 3.18 (d, *J*=1.6 Hz, 1H), 3.13 (q, *J*=7.0 Hz, 1H), 2.51 (m, 1H), 2.17 (s, 1H), 2.05–1.96 (m, 2H), 1.97–1.91 (m, 1H), 1.69 (ddd, *J*=14.0, 10.4, 1H), 1.55–1.48 (m, 5H), 1.11 (d, *J*=7.1 Hz, 3H), 1.06 (d, *J*=7.5 Hz, 3H), 1.03 (d, *J*=6.7 Hz, 3H) ppm
^{13}C-NMR (151 MHz, CDCl$_3$) δ = 170.93, 140.83, 140.18, 128.75 (2C), 128.37, 126.41 (2C), 122.99, 99.60, 84.83, 80.05, 76.30, 74.32, 72.39, 69.11, 65.78, 58.36, 57.49, 56.44, 46.38, 38.78, 35.80, 35.46, 33.53, 31.78, 31.08, 12.69, 11.71, 10.51 ppm
MS: m/z = 559 $[M+Na]^+$

(2*S*,5*R*,10*S*,11*R*,12*E*,14*S*,15*S*,16*S*,17*S*,18*R*)-6-chloro-1,17-dihydroxy-10,11,18-trimethoxy-2,14,16-trimethyl-5-phenyl-4,19-dioxabicyclo[13.3.1]nonadec-12-en-3-one (**1b**)



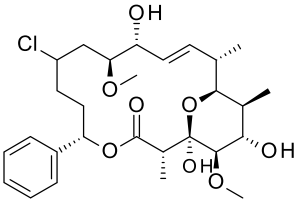
Isolated yield: 4% (4.1 mg)

^{1}H-NMR (600 MHz, $CDCl_3$) δ = 7.41–7.35 (m, 4H), 7.35–7.32 (m, 1H), 6.22 (dd, *J*=16.2, 3.9 Hz, 1H), 5.88 (d, *J*=9.1Hz, 1H), 5.48 (ddd, *J*=16.2, 9.5, 1.8 Hz, 1H), 4.39 (s, 1H), 4.38–4.35 (m, 1H), 4.12–4.06 (m, 1H), 3.79 (dd, *J*=10.5, 2.7 Hz, 1H), 3.63 (dd, *J*=9.5, 2.1 Hz, 1H), 3.47–3.43 (m, 4H), 3.37 (s, 3H), 3.30 (s, 3H), 3.16–3.14 (m, 1H), 3.14–3.09 (m, 2H), 2.56–2.46 (m, 1H), 2.17 (s, 1H) 2.03 (qd, *J*=9.9, 4.5 Hz, 1H) 1.90 (q, *J*=7.5 Hz, 1H) 1.87–1.80 (m, 1H), 1.66–1.61 (m, 2H), 1.45–1.36 (m, 1H), 1.35–1.25 (m, 2H), 1.07 (d, *J*=6.3 Hz, 3H), 1.06 (d, *J*=6.0 Hz, 3H), 1.02 (d, *J*=6.8 Hz, 3H) ppm;

^{13}C-NMR (151 MHz, $CDCl_3$) δ = 170.01, 140.07, 137.28, 128.99, 128.40 (2C), 128.05 (2C), 122.64, 99.66, 84.90, 83.23, 76.26, 75.65, 72.68, 69.27, 62.15, 58.25, 57.52, 56.39, 46.37, 35.82, 35.30, 33.83, 29.53, 18.98, 12.61, 11.55, 10.39 ppm.

HPLC-MS: rt = 1.80 min, m/z = 564/566 $[M+H+Na]^{2+}$, 539/541 $[M–H]^{–}$ chloro isotopic pattern

(1*R*,2*S*,5*S*,10*S*,11*R*,12*E*,14*S*,15*S*,16*S*,17*S*,18*R*)-8-chloro-1,11,17-trihydroxy-10,18-dimethoxy-2,14,16-trimethyl-5-phenyl-4,19-dioxabicyclo[13.3.1]nonadec-12-en-3-one (**2a**)



Isolated yield: 10% (16.1 mg)

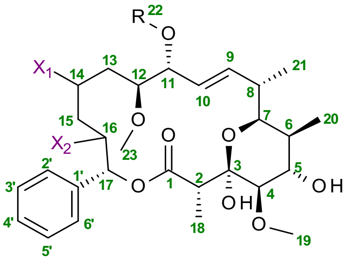
^{1}H-NMR (600 MHz, $CDCl_3$) δ = 7.38–7.33 (m, 4H), 7.33–7.28 (m, 1H), 6.16 (dd, *J*=16.1, 3.8 Hz, 1H), 5.89 (dd, *J*=11.7, 2.7 Hz, 1H), 5.50 (ddd, *J*=16.1, 9.4, 1.8 Hz, 1H), 4.58 (s, 1H), 4.19 (dd, *J*=9.3, 2.1 Hz, 1H), 4.08 (br s, 1H), 4.00 (tt, *J*=11.5, 2.7 Hz, 1H), 3.81 (dt, *J*=11.0, 2.9 Hz, 1H), 3.76 (dd, *J*=10.6, 2.6 Hz, 1H), 3.46 (s, 3H), 3.38 (s, 3H), 3.30 (br d, *J*=7.3 Hz, 1H), 3.17 (d, *J*=1.8 Hz, 1H), 3.12 (q, *J*=7.2 Hz, 1H), 2.51–2.43 (m, 1H), 2.39–2.29 (m, 1H), 2.17 (ddt, *J*=15.0, 11.6, 3.5, 3.5 Hz, 1H), 2.08–2.01 (m, 2H), 1.93–1.88 (m, 1H), 1.84 (ddt, *J*=14.8, 11.8, 3.4, 3.4 Hz, 1H), 1.75 (ddt, *J*=14.9, 11.2, 4.0, 4.0 Hz, 1H), 1.69–1.64 (m, 2H), 1.11 (d, *J*=7.0 Hz, 3H), 1.06 (d, *J*=7.5 Hz, 3H), 1.00 (d, *J*=6.8 Hz, 3H) ppm

^{13}C-NMR (151 MHz, $CDCl_3$) δ = 170.83, 140.85, 137.99, 128.77 (2C), 128.34, 126.17 (2C), 125.19, 99.54, 81.27, 76.29, 74.54, 73.31, 72.42, 69.18, 58.09, 57.47, 56.58, 46.39, 39.15, 35.86, 35.35, 34.29, 32.84, 12.60, 11.71, 10.47 ppm

MS: m/z = 563/565 $[M+Na]^{+}$ chloro isotopic pattern

Supplementary Table 12. Assignment Table ^{1}H-NMR.

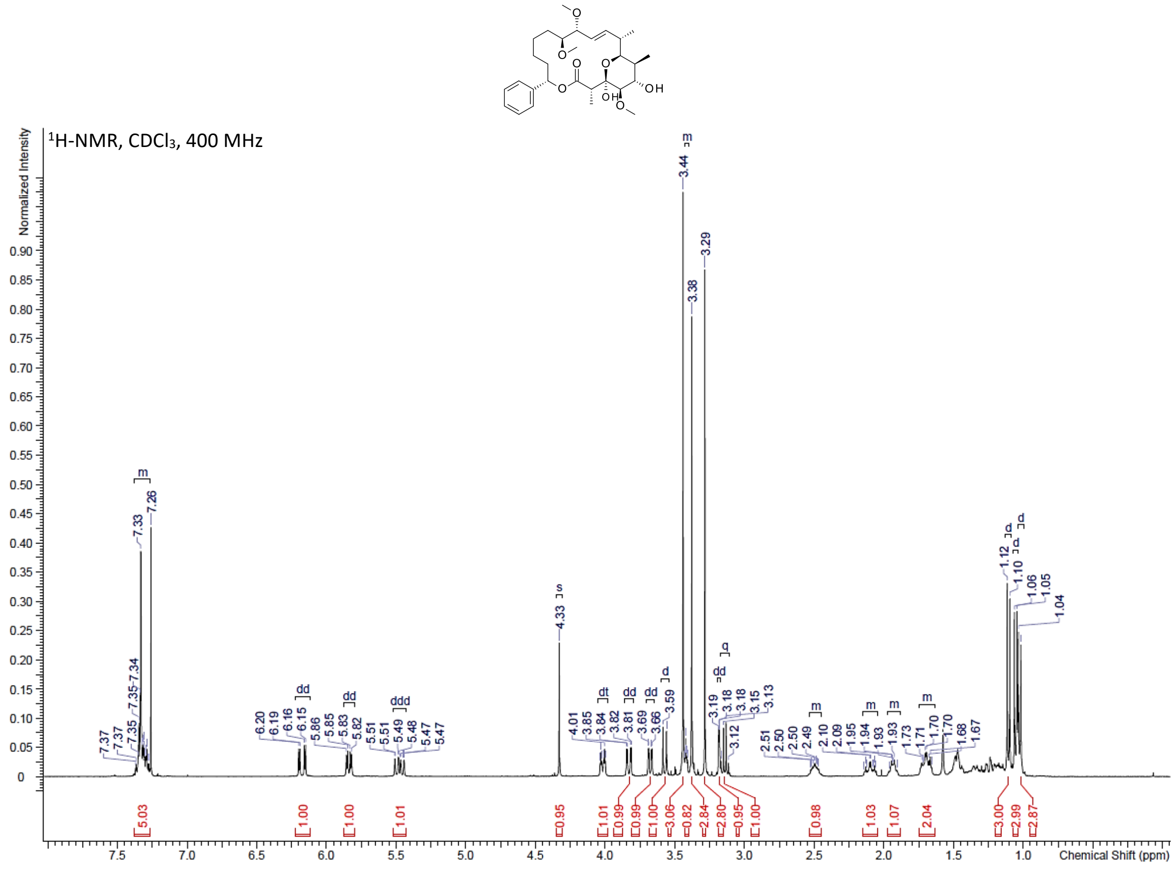
Chemical shifts (in ppm) for multiplets were taken at the center of the multiplet for clarity.
The assignment of the peaks was done based on analysis of 2D spectra: COSY, HSQC, HMBC, ROESY.



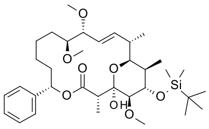
Carbon number	Soraphen A	(1a) R=Me; X_1=Cl, X_2=H	(1b) R=Me; X_1=H, X_2=Cl	(1c) R=Me; X_1=OH, X_2=H	Soraphen C R=H; X_1=H, X_2=H	(2a) R=H; X_1=Cl, X_2=H
2	3.14	3.15	3.12	3.16	3.15	3.12
4	3.18	3.19	3.15	3.20	3.18	3.17
5	4.02	4.12	4.09	4.07	4.02	4.00
6	1.94	1.95	1.90	1.95	1.93	1.90
7	3.83	3.79	3.79	3.84	3.82	3.76
8	2.49	2.54	2.51	2.54	2.45	2.47
9	6.17	6.23	6.22	6.19	6.15	6.16
10	5.48	5.51	5.48	5.51	5.49	5.50
11	3.68	3.71	3.63	3.75	4.17	4.19
12	3.41	3.91	3.45	3.77	3.36	3.81
13	1.24	1.72	1.34	1.52	1.17	1.66
	1.69	2.01	1.64	1.71	1.77	2.05
14	1.16	4.02	1.43	3.86	1.17	4.00
	1.46		1.64		1.51	
15	1.34	1.77	1.86	1.58	1.35	1.75
	1.46	1.89	2.03	1.58	1.48	1.85
16	1.67	2.07	4.37	2.02	1.67	2.05
	2.10	2.20		2.05	2.10	2.18
17	5.84	5.95	5.88	5.90	5.82	5.89
18	1.11	1.15	1.07	1.14	1.09	1.11
19	3.38	3.41	3.37	3.41	3.38	3.38
20	1.05	1.10	1.06	1.19	1.05	1.06
21	1.03	1.05	1.02	1.05	1.00	1.00
22	3.29	3.35	3.30	3.33	/	/
23	3.44	3.50	3.45	3.49	3.43	3.46
2’	7.32	7.38	7.38	7.37	7.35	7.35
3’	7.32	7.38	7.38	7.38	7.35	7.35
4’	7.32	7.32	7.34	7.33	7.27	7.30
5’	7.32	7.38	7.38	7.38	7.35	7.35
6’	7.32	7.38	7.38	7.37	7.35	7.35

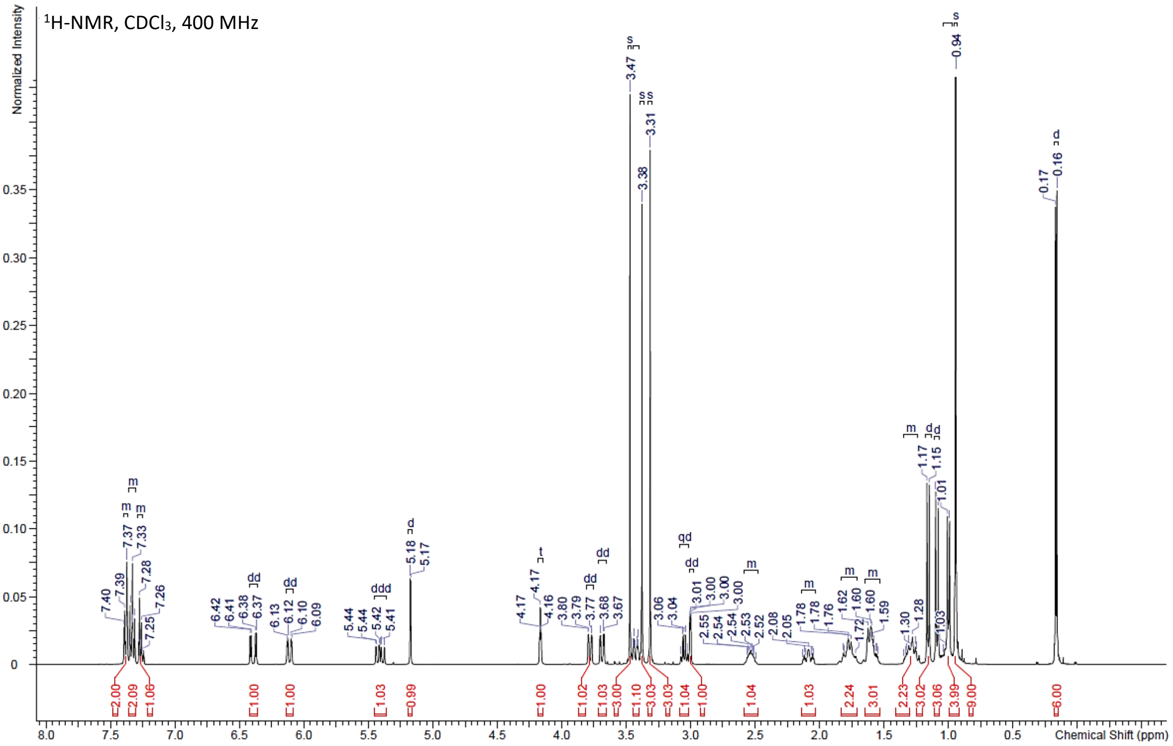
NMR spectra

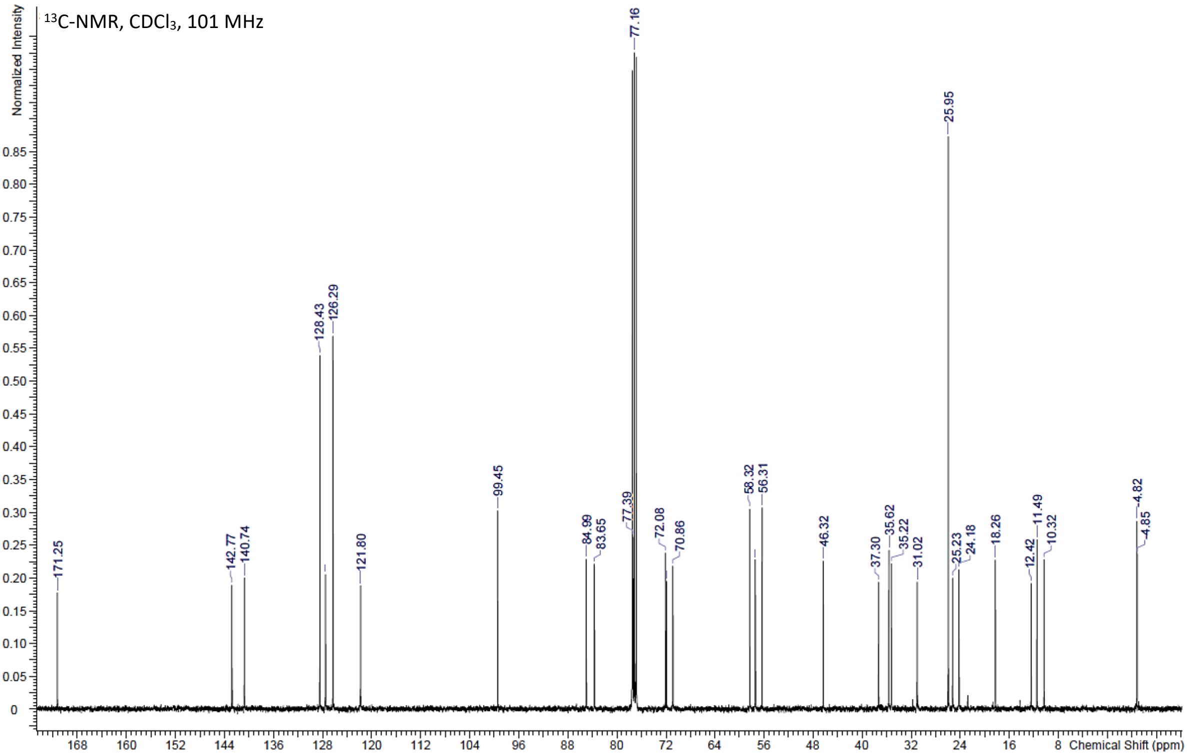
Supplementary Figure 16. NMR of compound **1**.



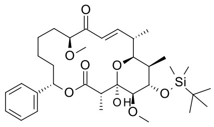
Supplementary Figure 17. NMR of compound **S1**.

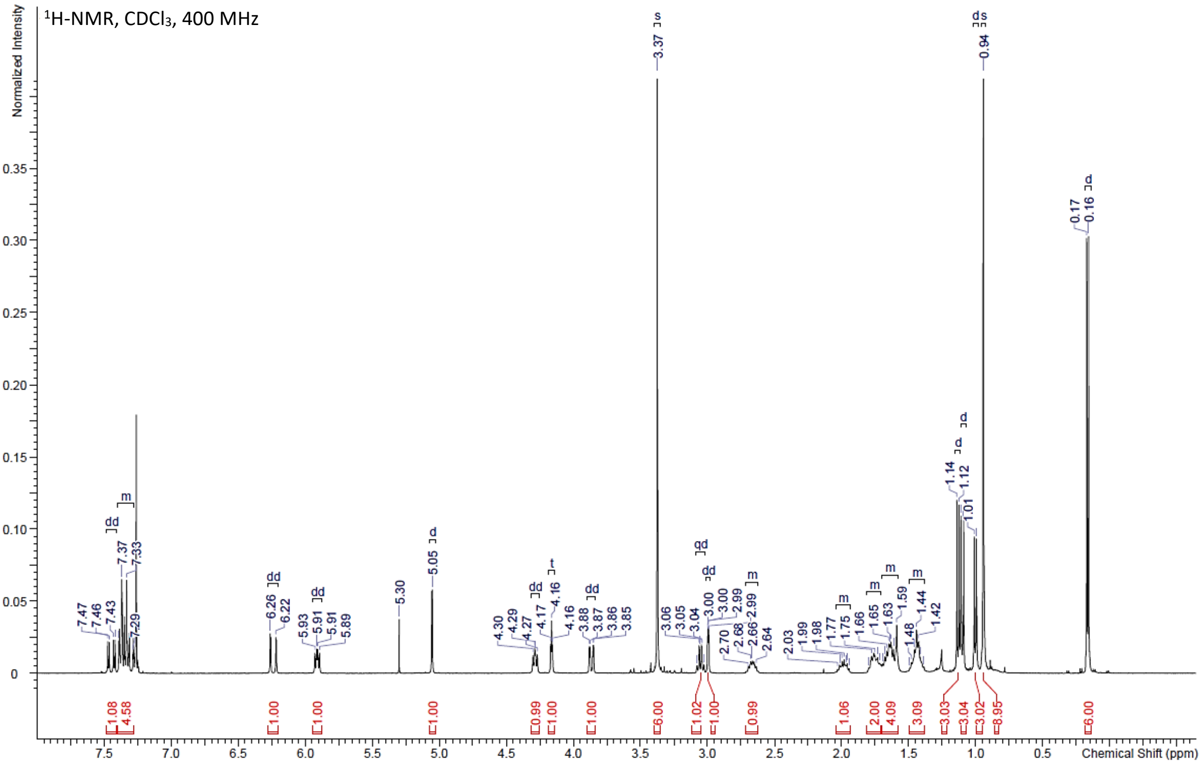


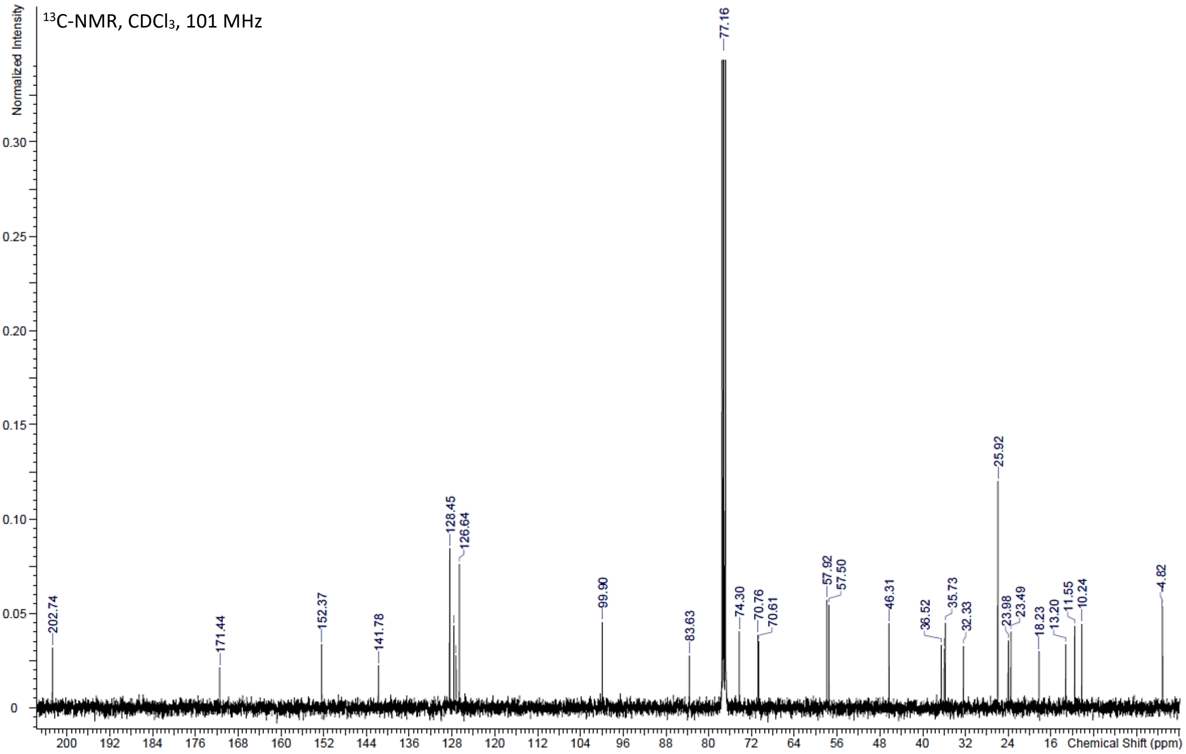




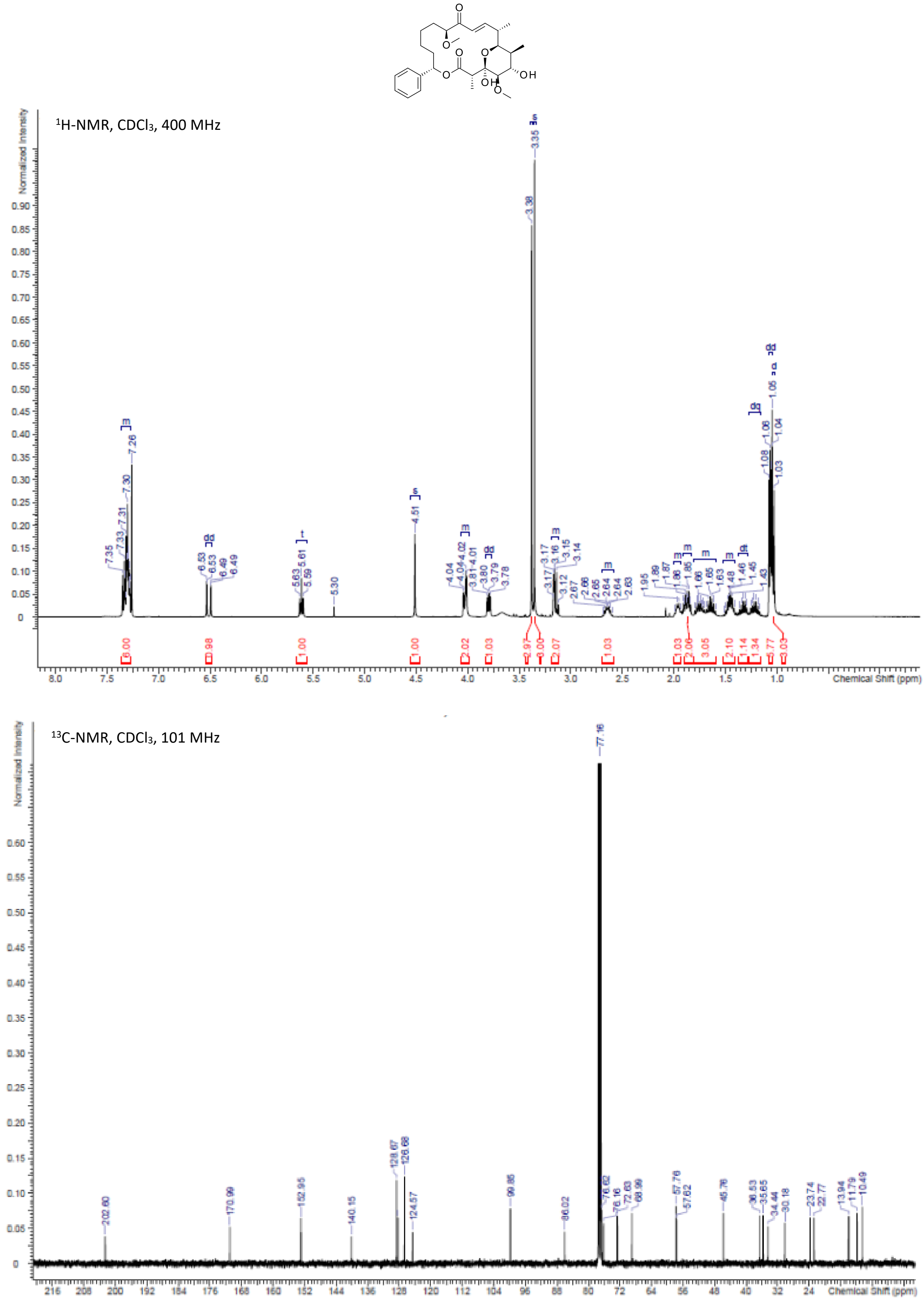
Supplementary Figure 18. NMR of compound **S2**.



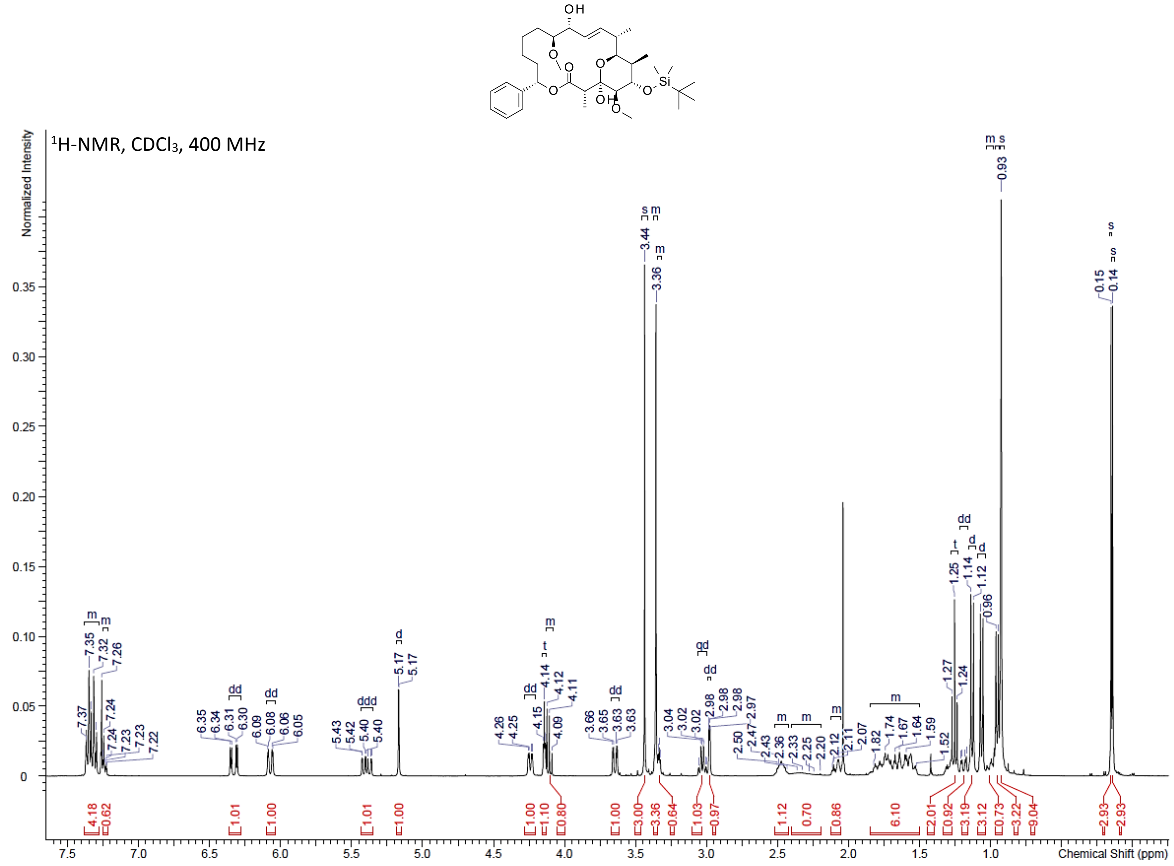

^{1}H-NMR, $CDCl_3$, 400 MHz


^{13}C-NMR, $CDCl_3$, 101 MHz

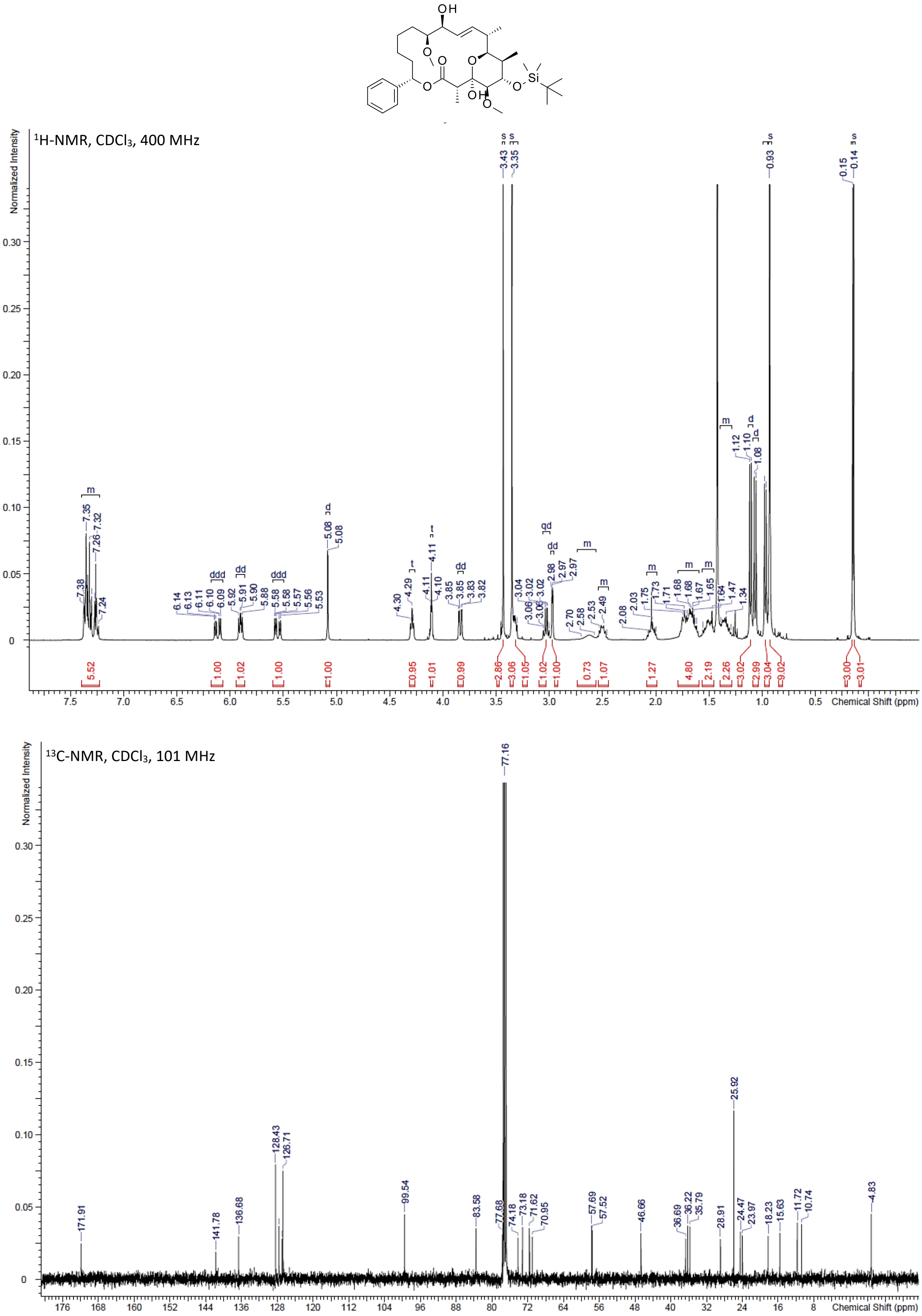
Supplementary Figure 19. NMR of compound **3**.



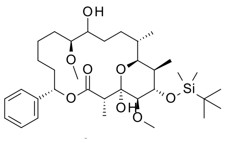
Supplementary Figure 20. NMR of compound **S3**.

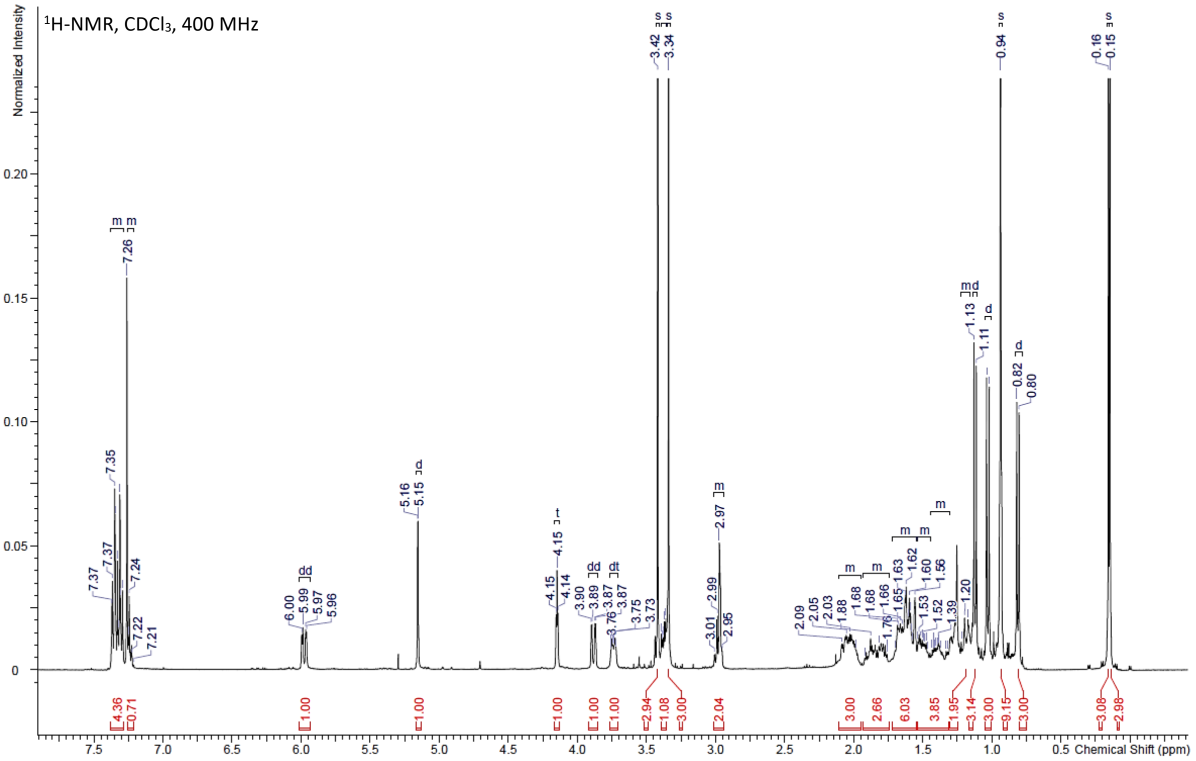


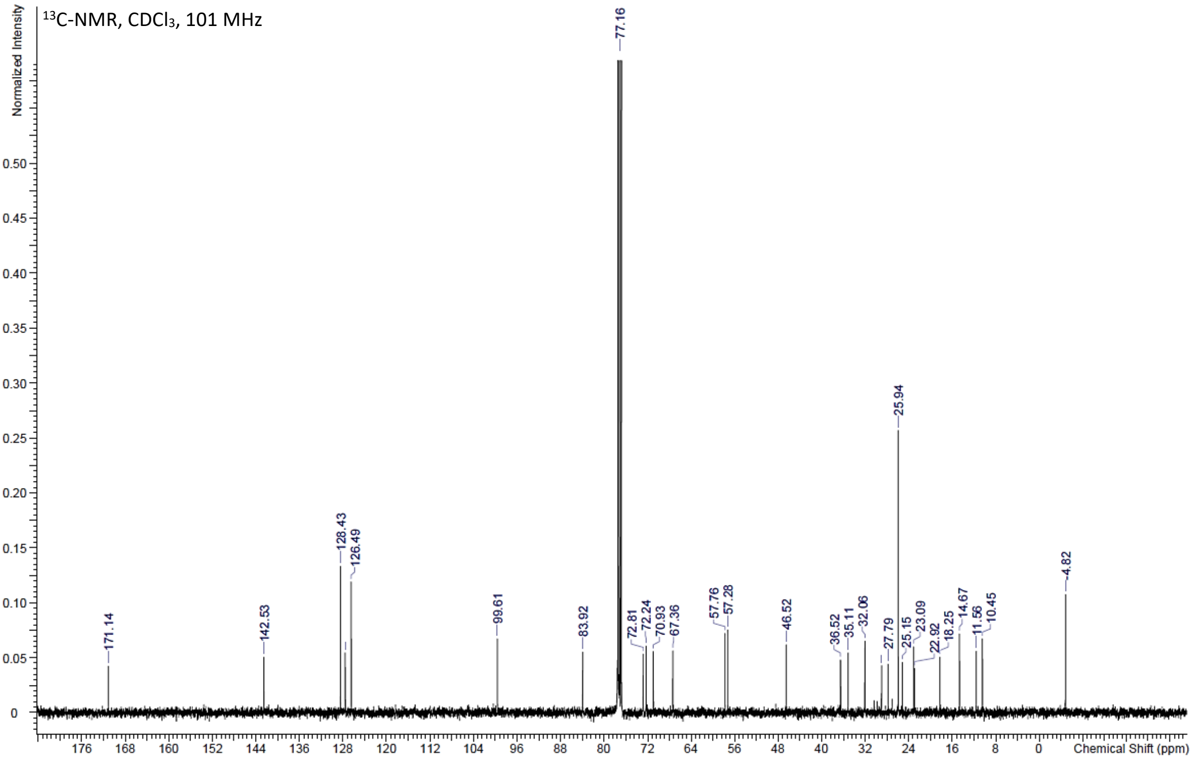
Supplementary Figure 21. NMR of compound **S4**.



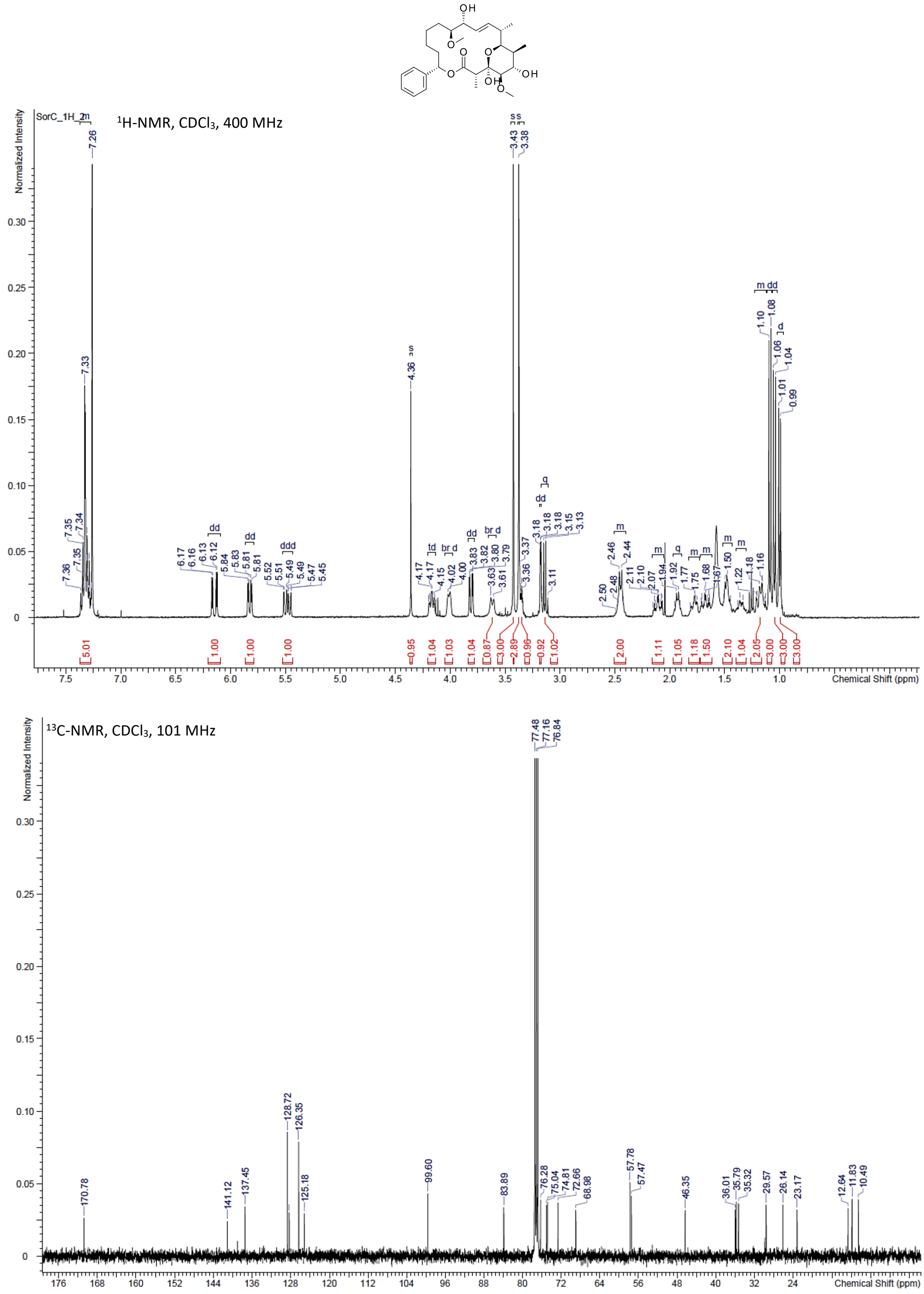
Supplementary Figure 22. NMR of compound **S5**.



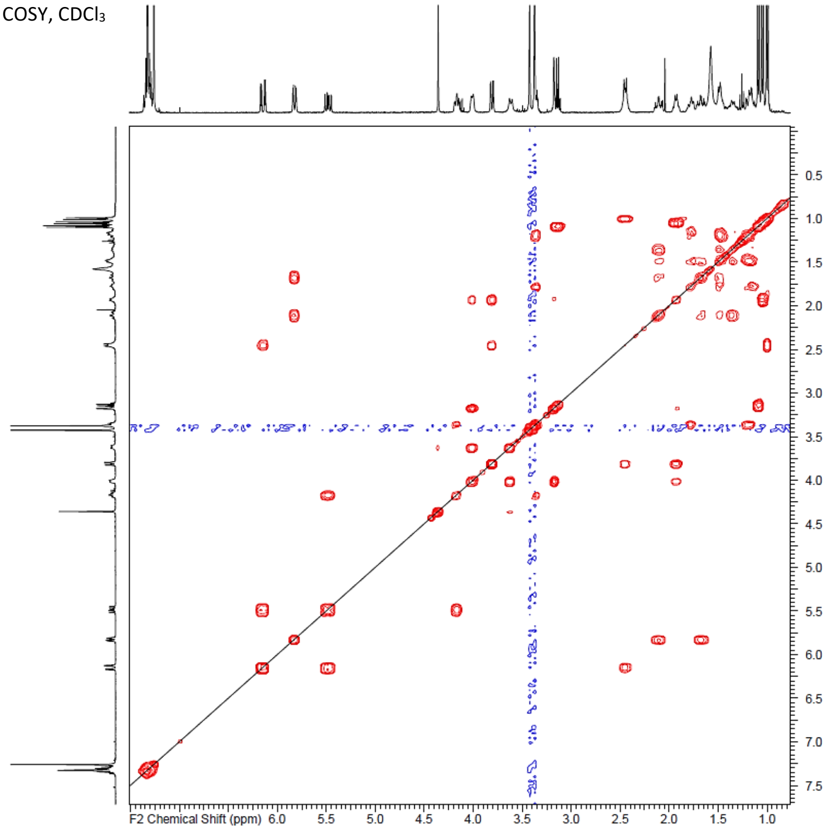




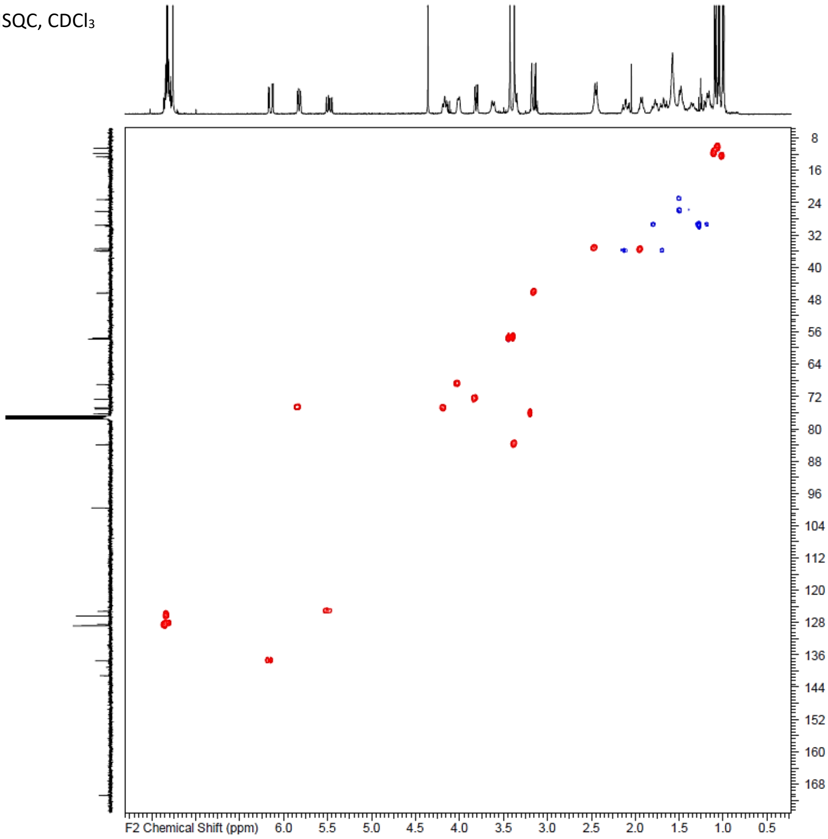
Supplementary Figure 23. NMR of compound soraphen C, **2**.



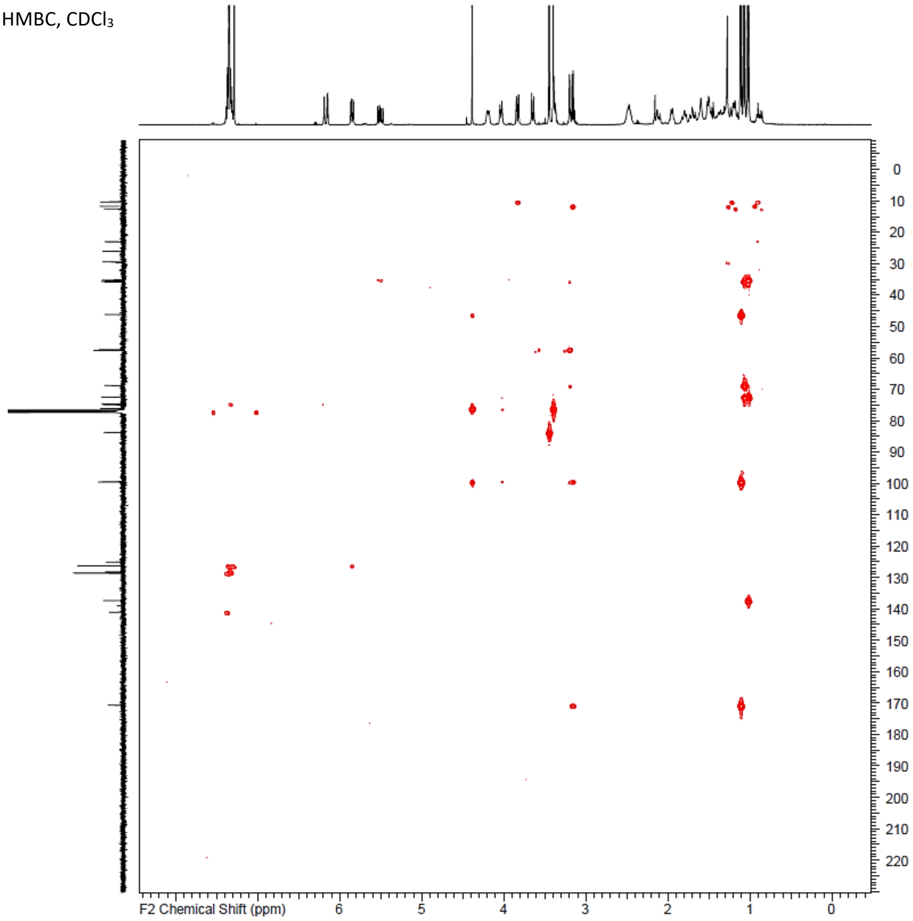
COSY, $CDCl_3$



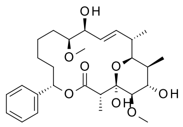
HSQC, $CDCl_3$

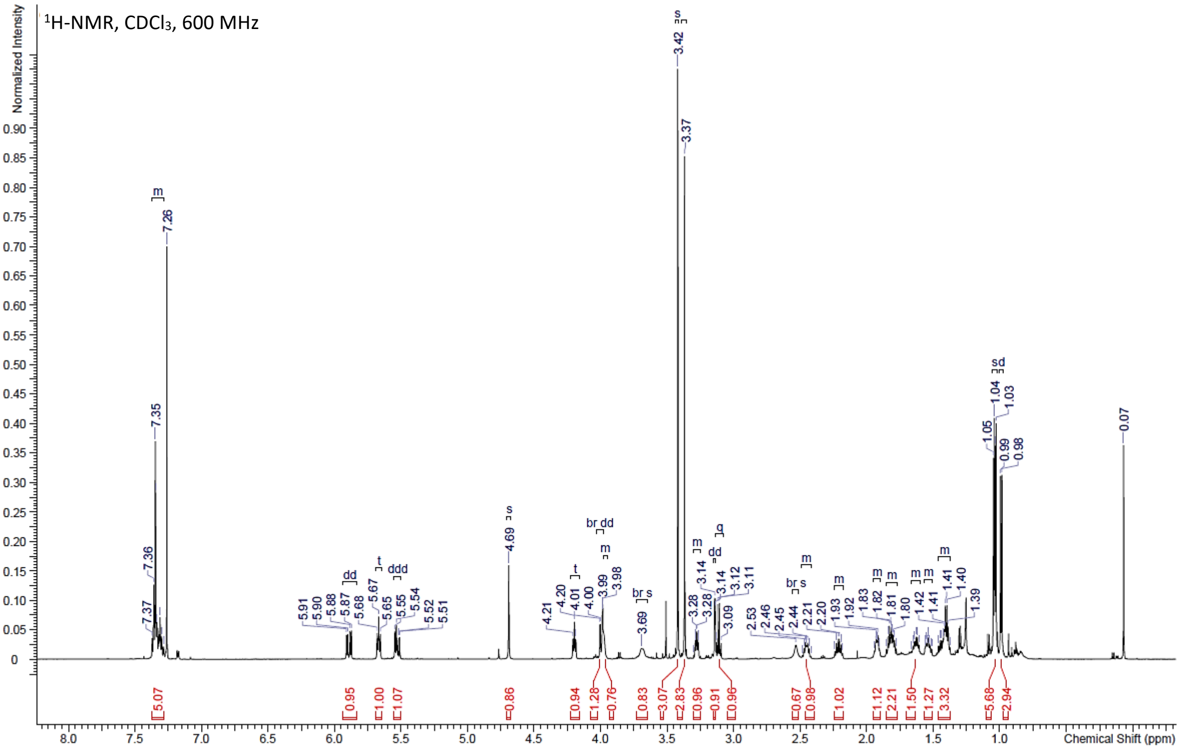


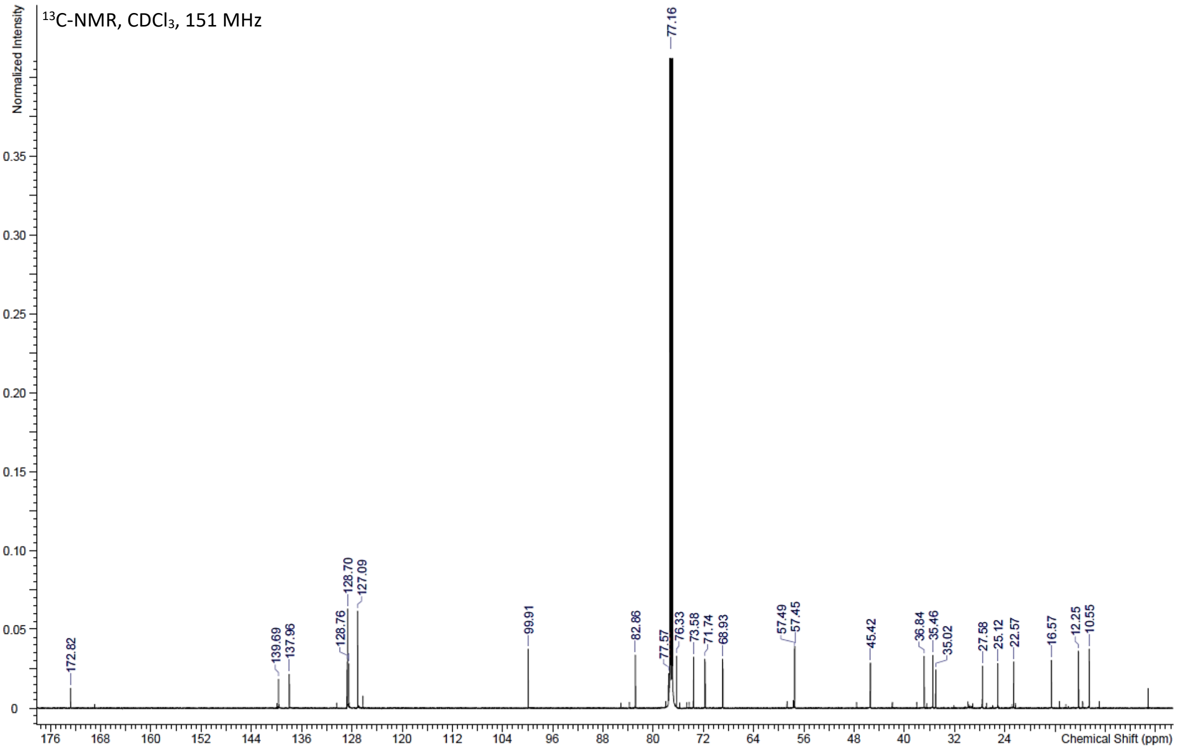
HMBC, CDCl$_3$


F2 Chemical Shift (ppm)
6
5
4
3
2
1
0
0
10
20
30
40
50
60
70
80
90
100
110
120
130
140
150
160
170
180
190
200
210
220

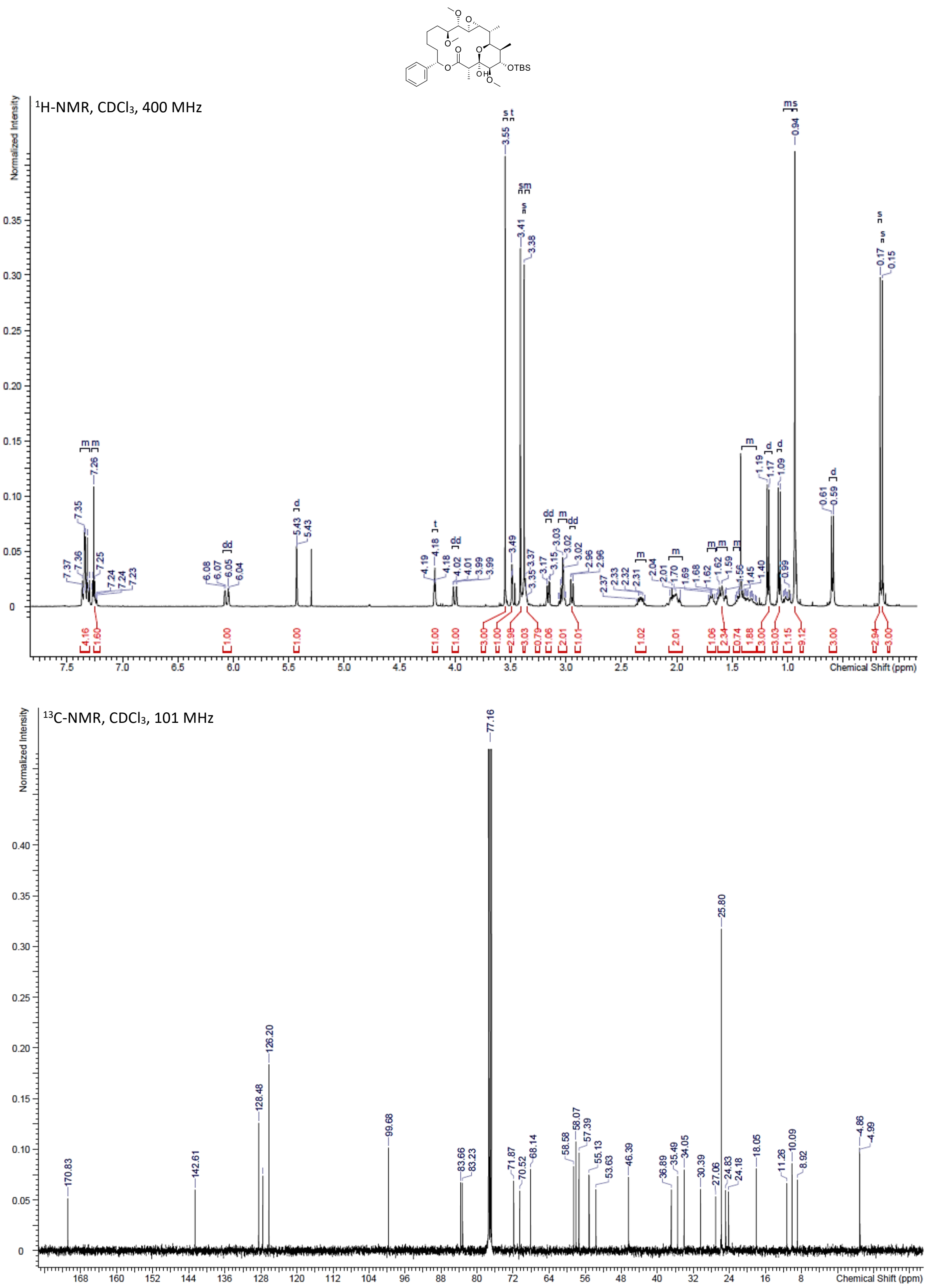
Supplementary Figure 24. NMR of compound *epi*-soraphen C, **S7**.



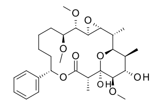


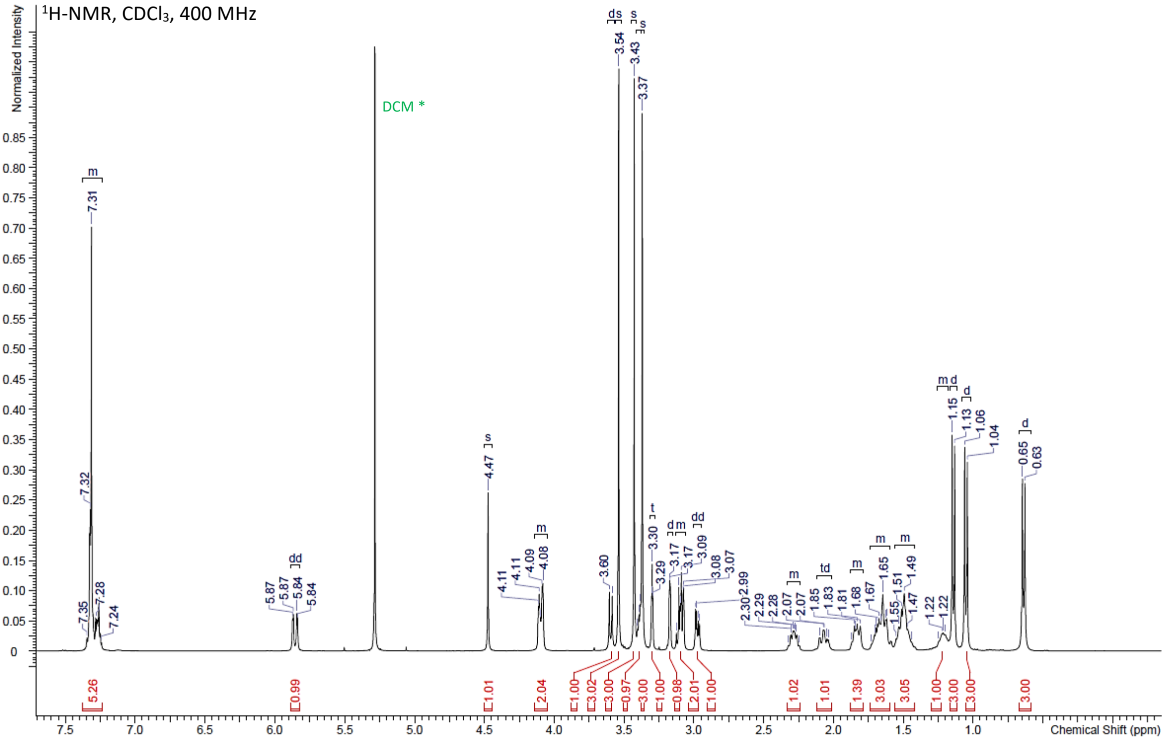


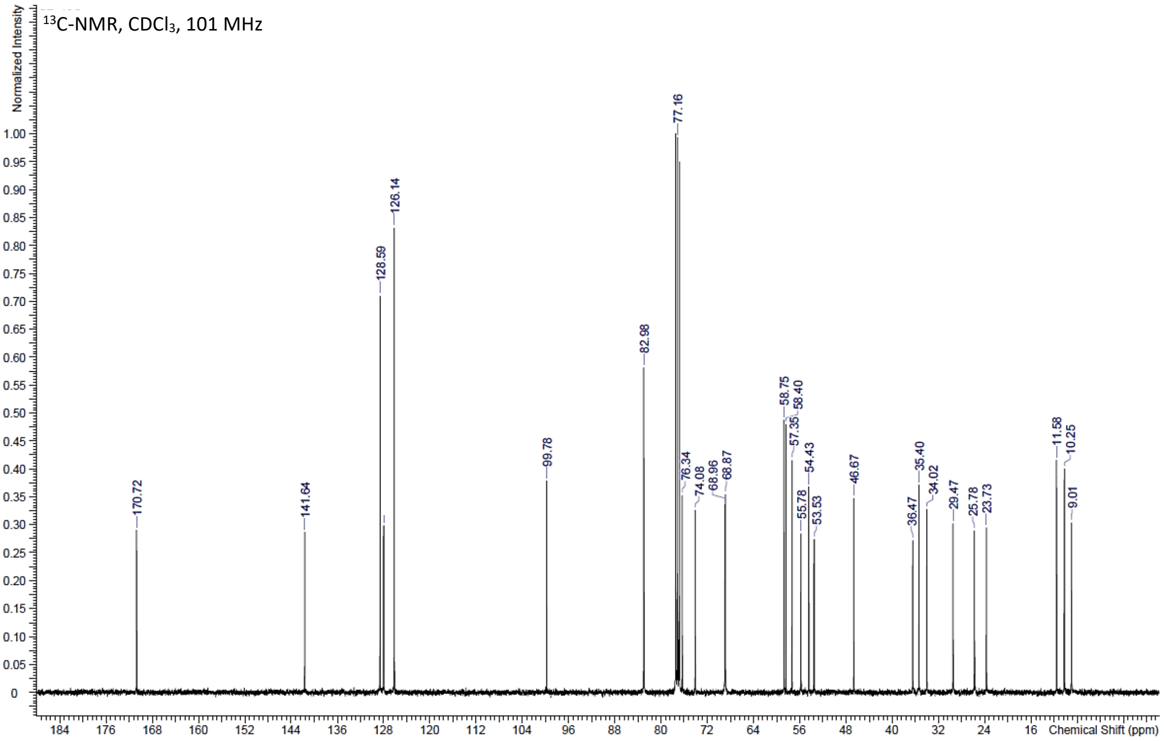
Supplementary Figure 25. NMR of compound **S6**.



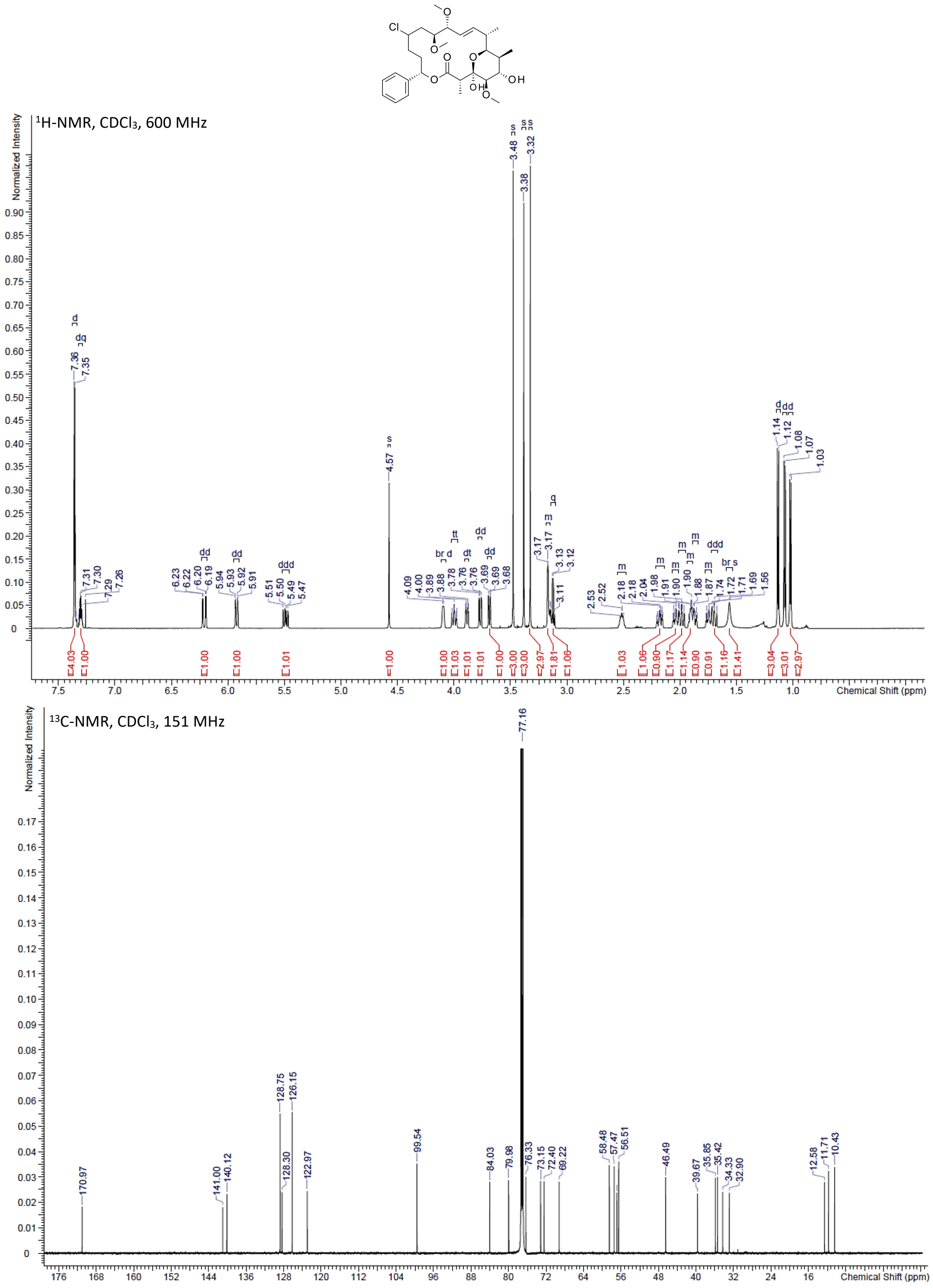
Supplementary Figure 26. NMR of compound **4**.



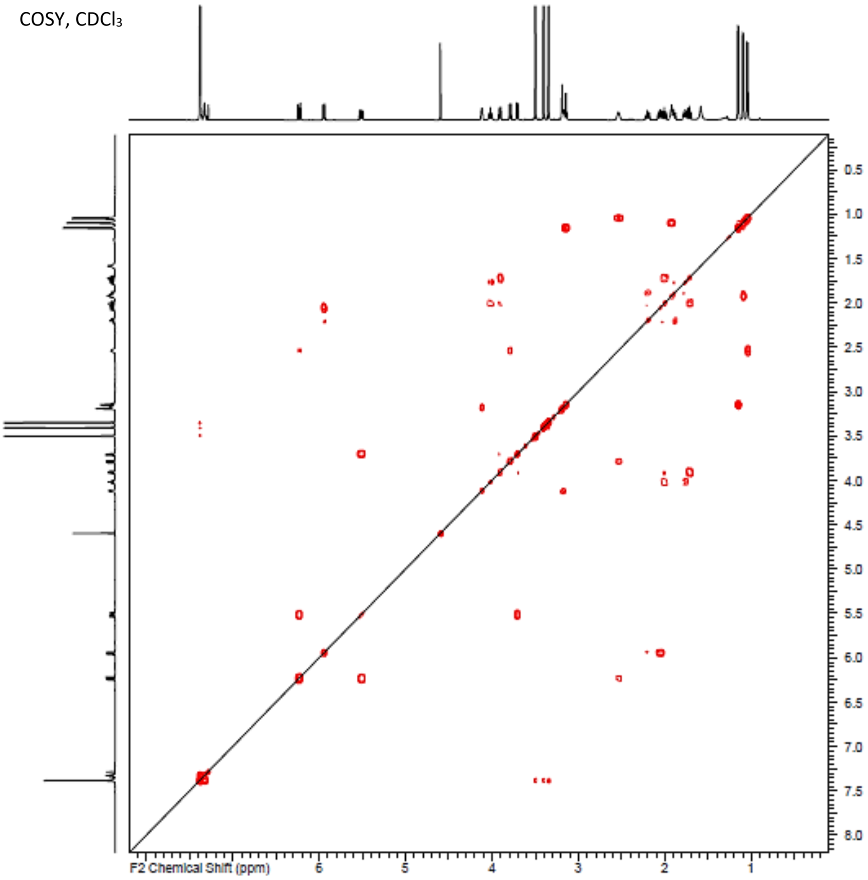




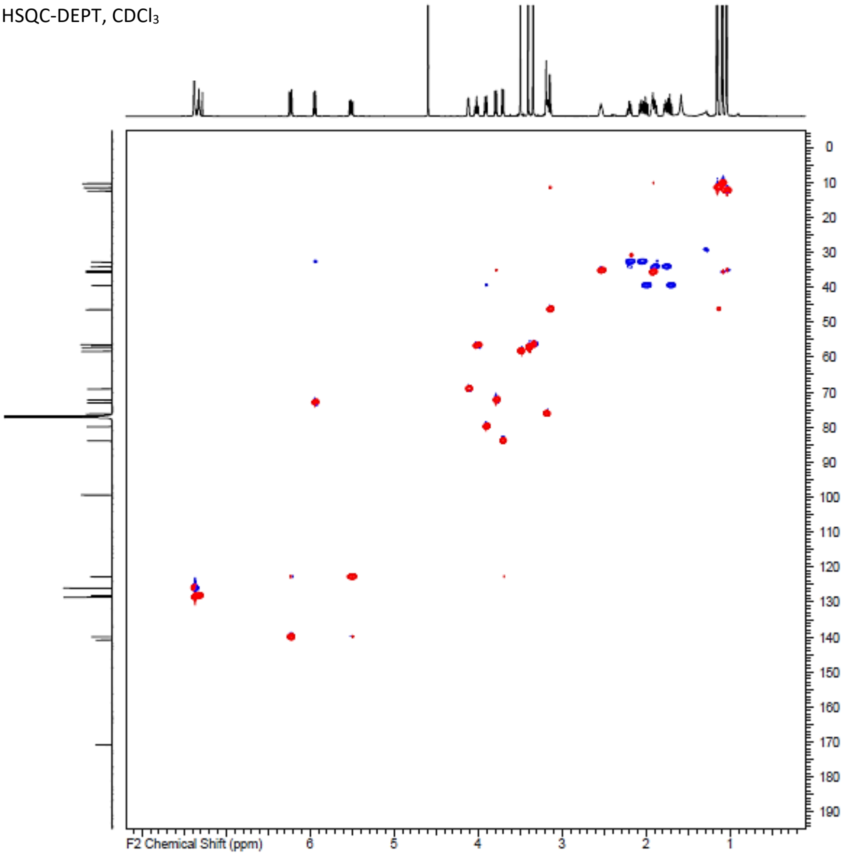
Supplementary Figure 27. NMR of compound **1a**.



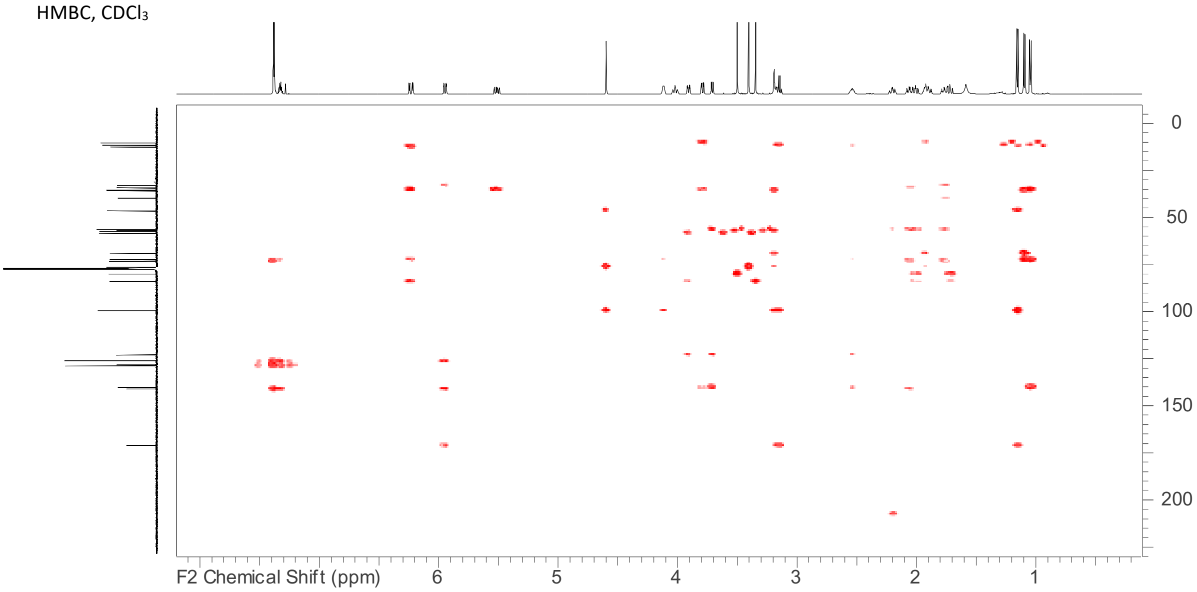
COSY, $CDCl_3$



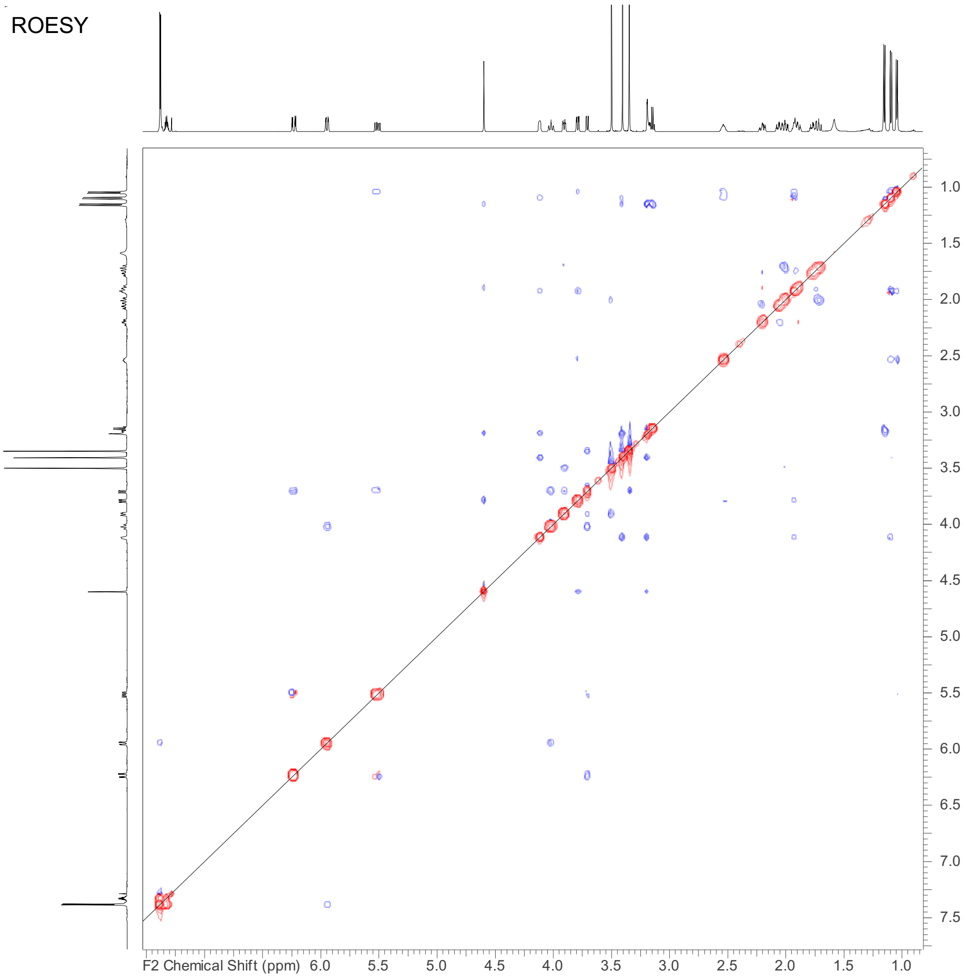
HSQC-DEPT, $CDCl_3$



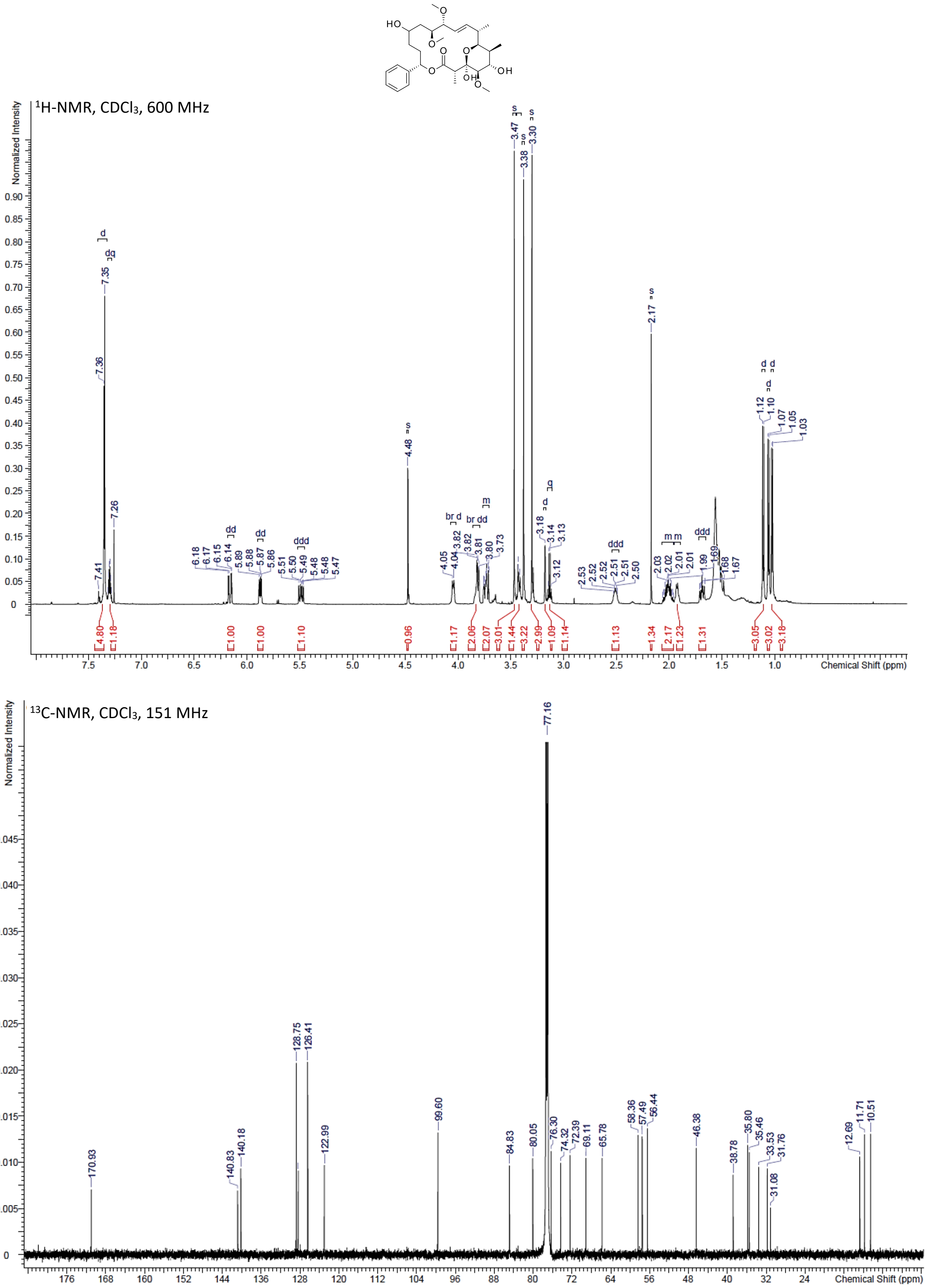
HMBC, $CDCl_3$



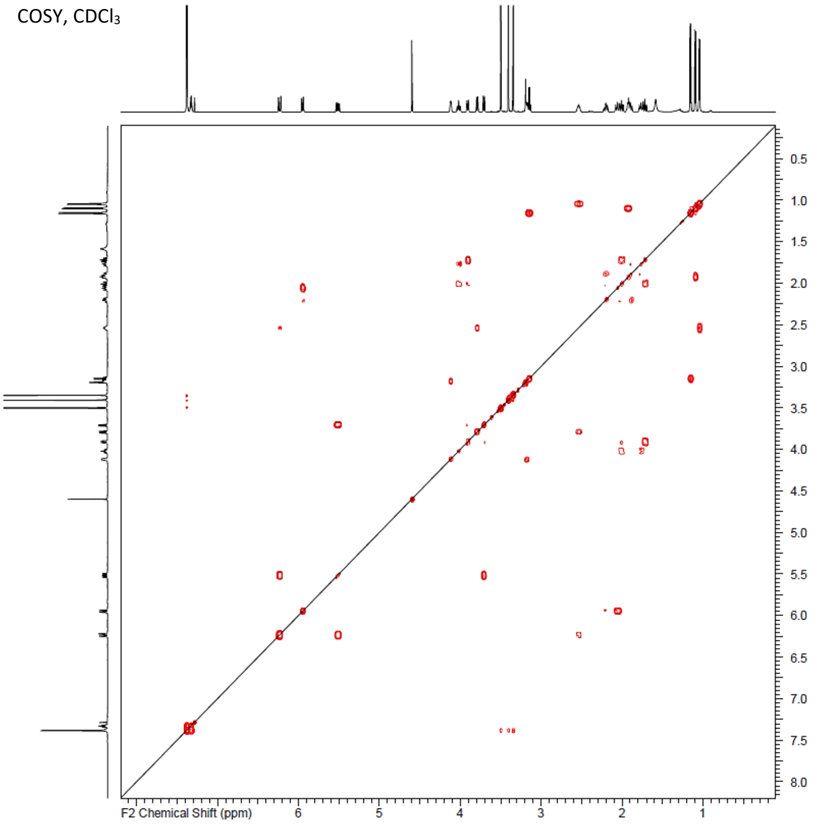
ROESY



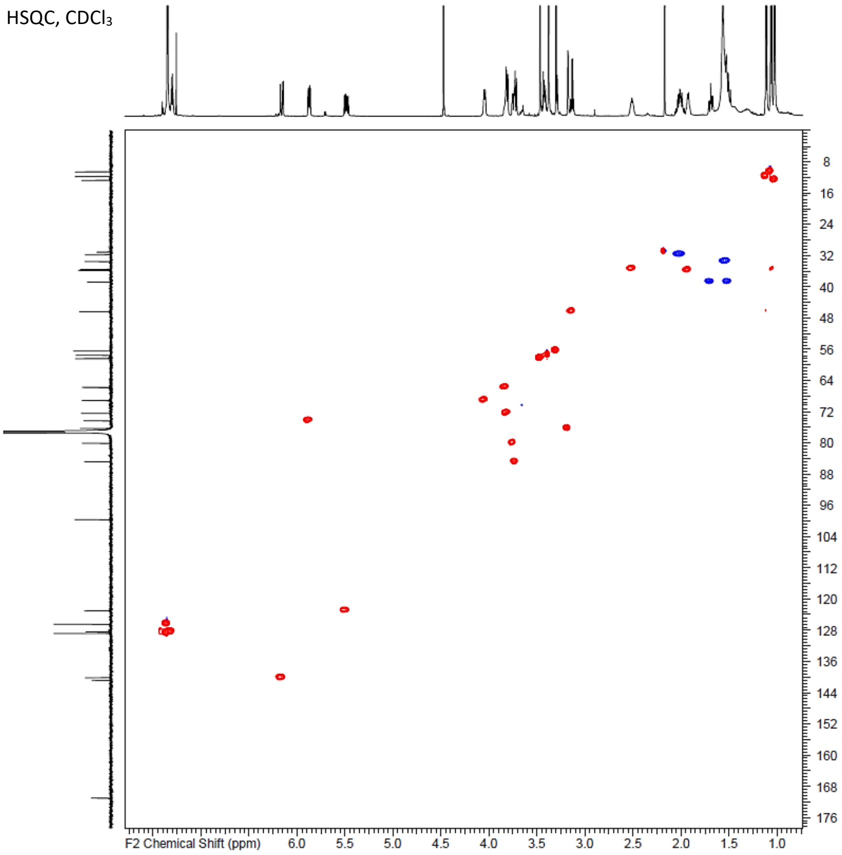
Supplementary Figure 28. NMR of compound **1c**.



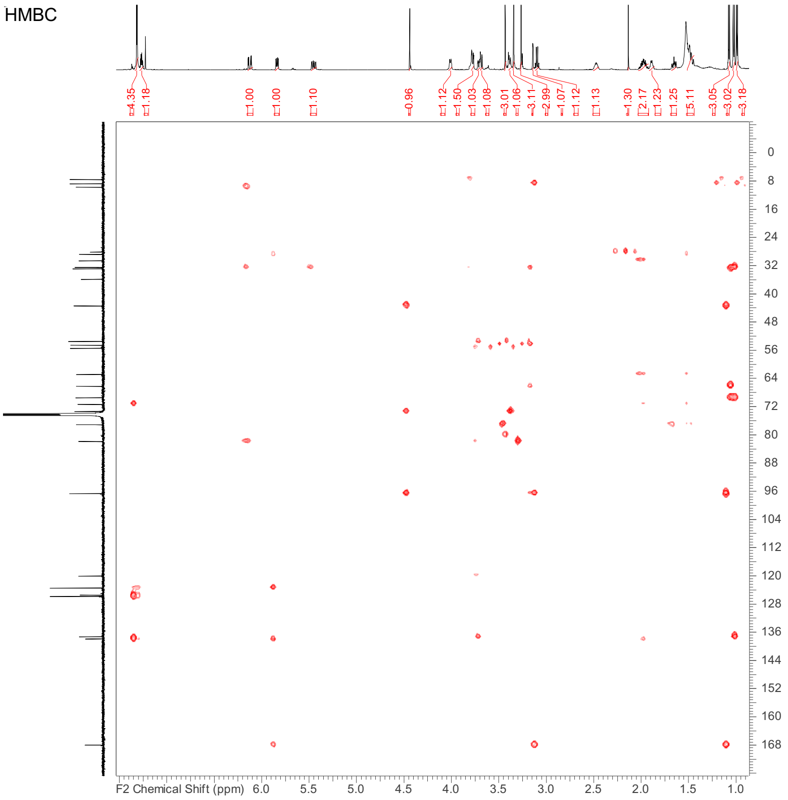
COSY, $CDCl_3$



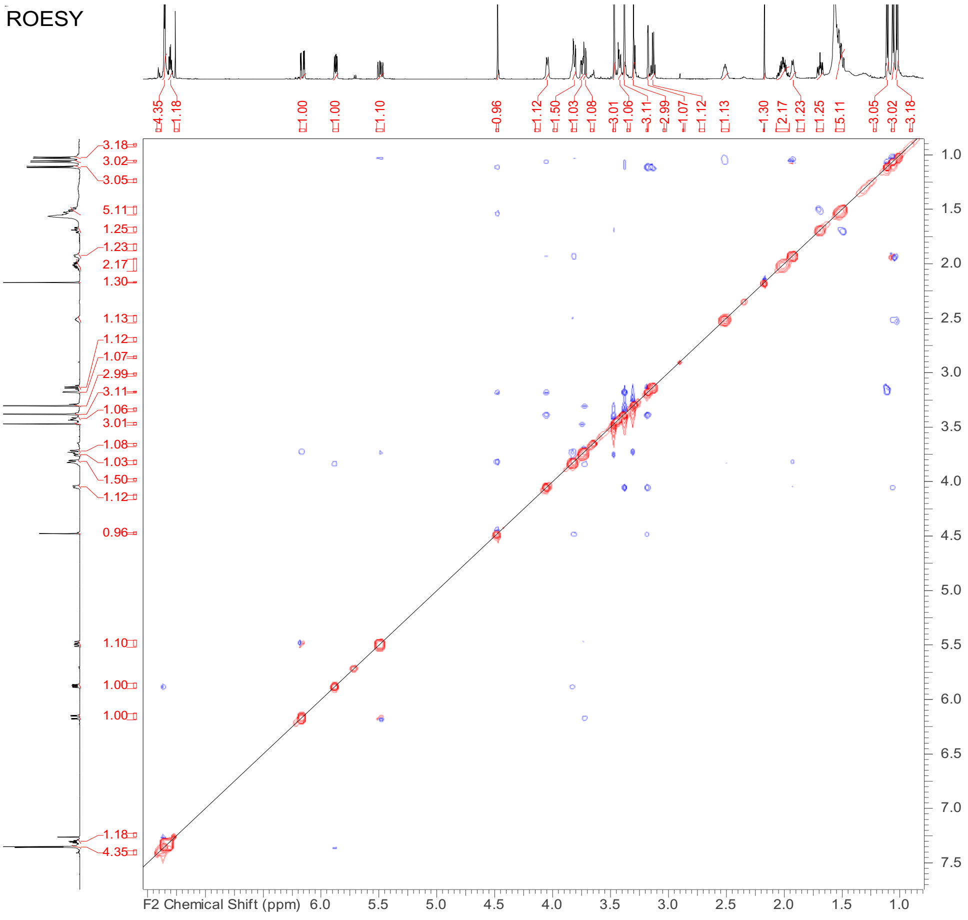
HSQC, $CDCl_3$



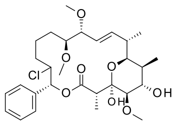
HMBC

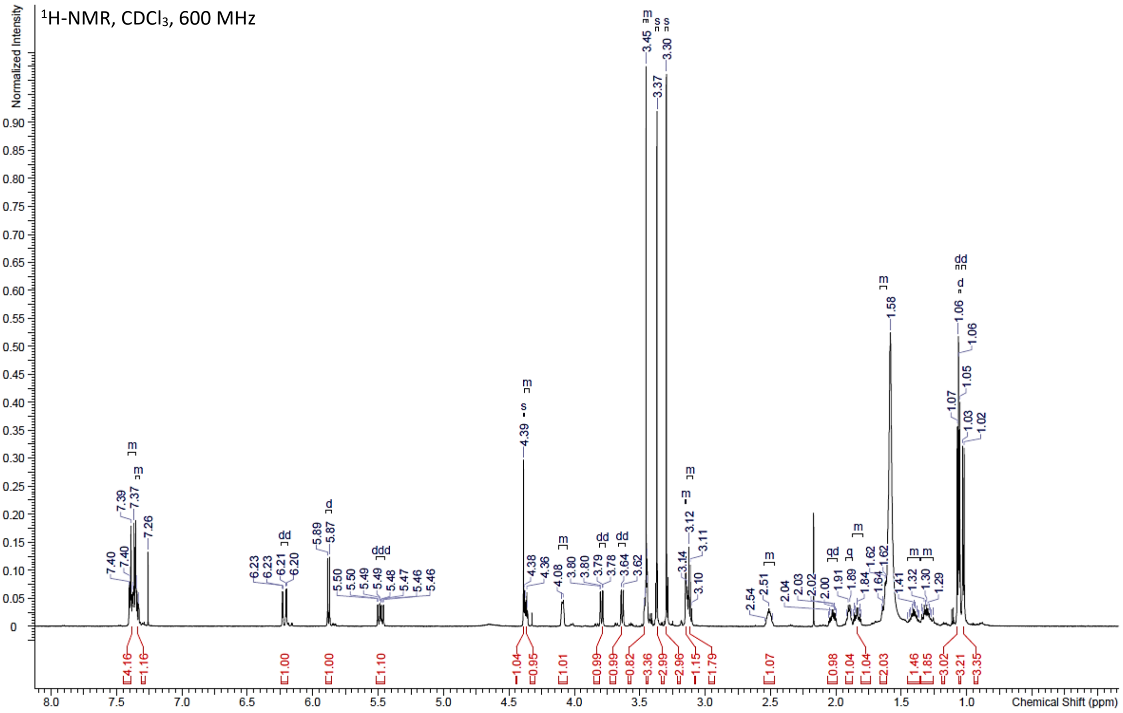

4.35
1.18
1.00
1.00
1.10
0.96
1.12
1.50
1.03
1.08
3.01
1.06
3.11
2.99
1.07
1.12
1.13
1.30
2.17
1.23
1.25
5.11
3.05
3.02
3.18
0
8
16
24
32
40
48
56
64
72
80
88
96
104
112
120
128
136
144
152
160
168
F2 Chemical Shift (ppm) 6.0
5.5
5.0
4.5
4.0
3.5
3.0
2.5
2.0
1.5
1.0

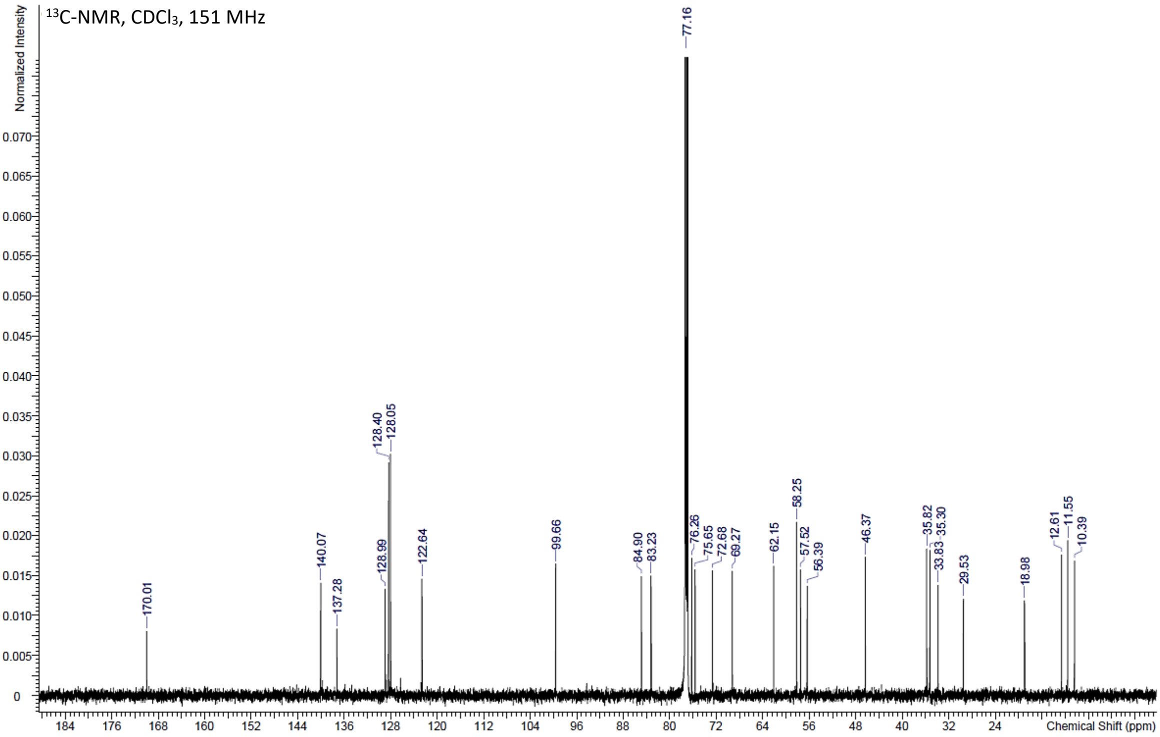
ROESY


4.35
1.18
1.00
1.00
1.10
0.96
1.12
1.50
1.03
1.08
3.01
1.06
3.11
2.99
1.07
1.12
1.13
1.30
2.17
1.23
1.25
5.11
3.05
3.02
3.18
1.0
1.5
2.0
2.5
3.0
3.5
4.0
4.5
5.0
5.5
6.0
6.5
7.0
7.5
F2 Chemical Shift (ppm) 6.0
5.5
5.0
4.5
4.0
3.5
3.0
2.5
2.0
1.5
1.0

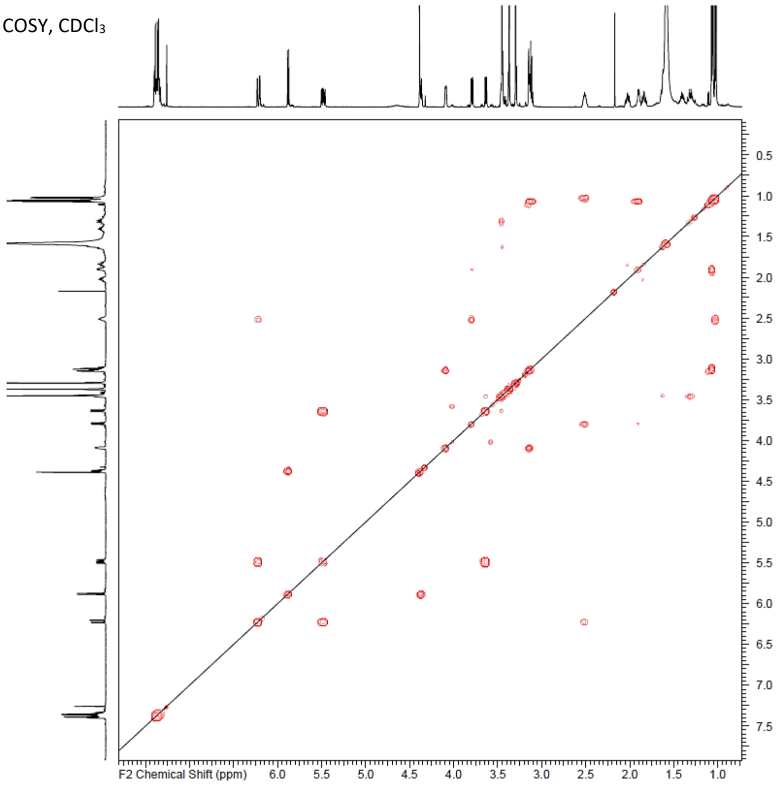
Supplementary Figure 29. NMR of compound **1b**.



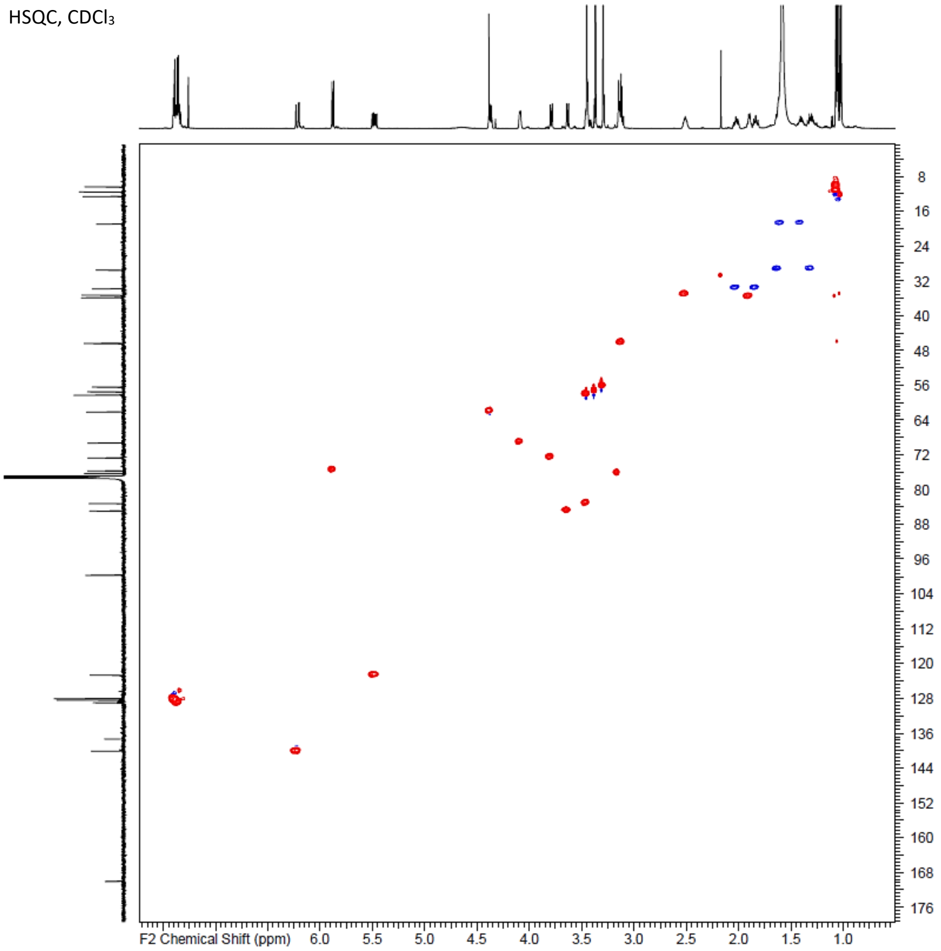

^{1}H-NMR, CDCl$_3$, 600 MHz


^{13}C-NMR, CDCl$_3$, 151 MHz

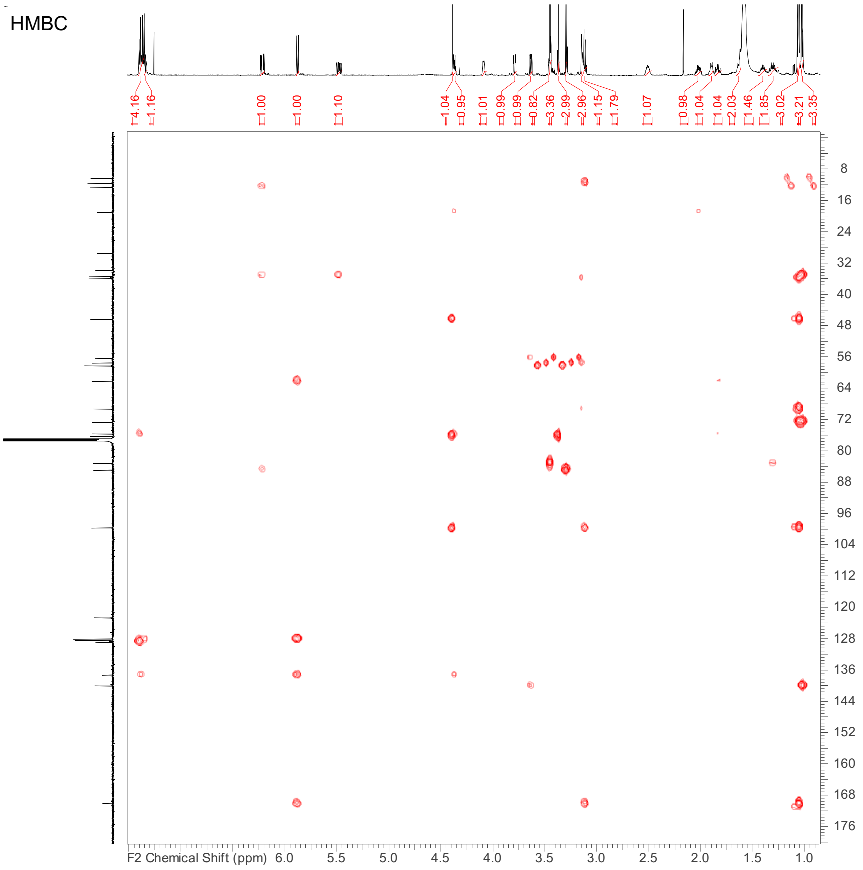
COSY, $CDCl_3$



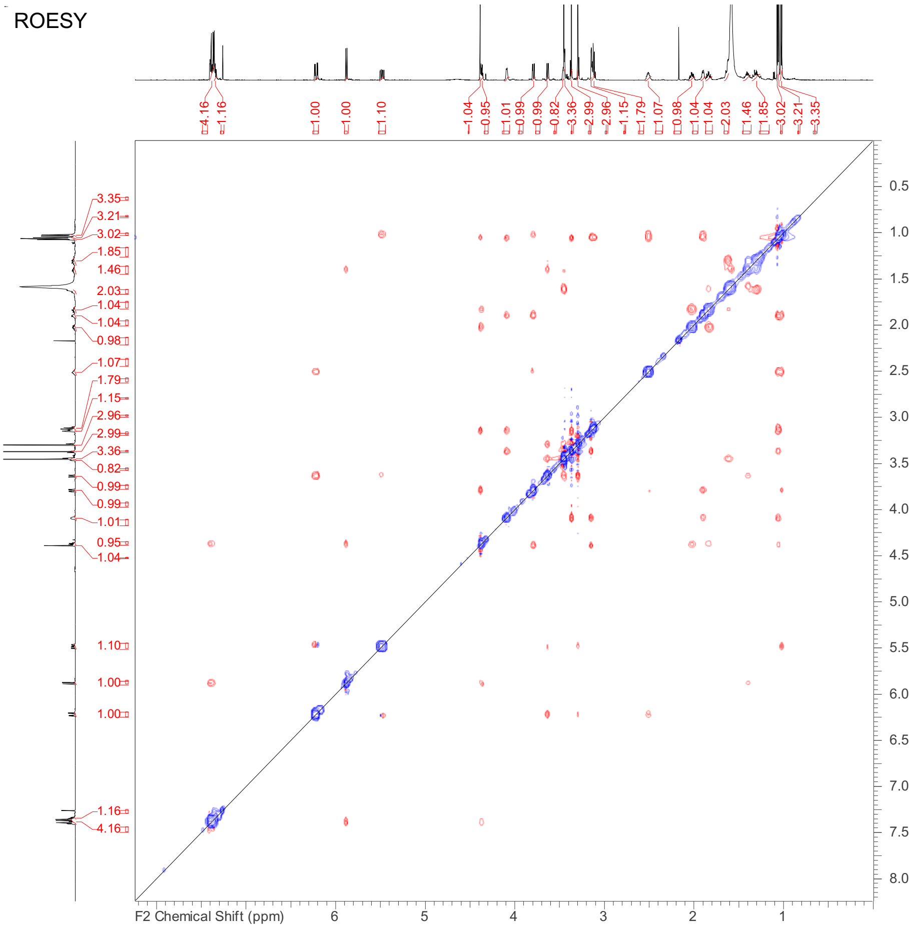
HSQC, $CDCl_3$



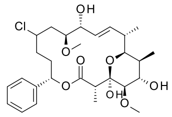
HMBC

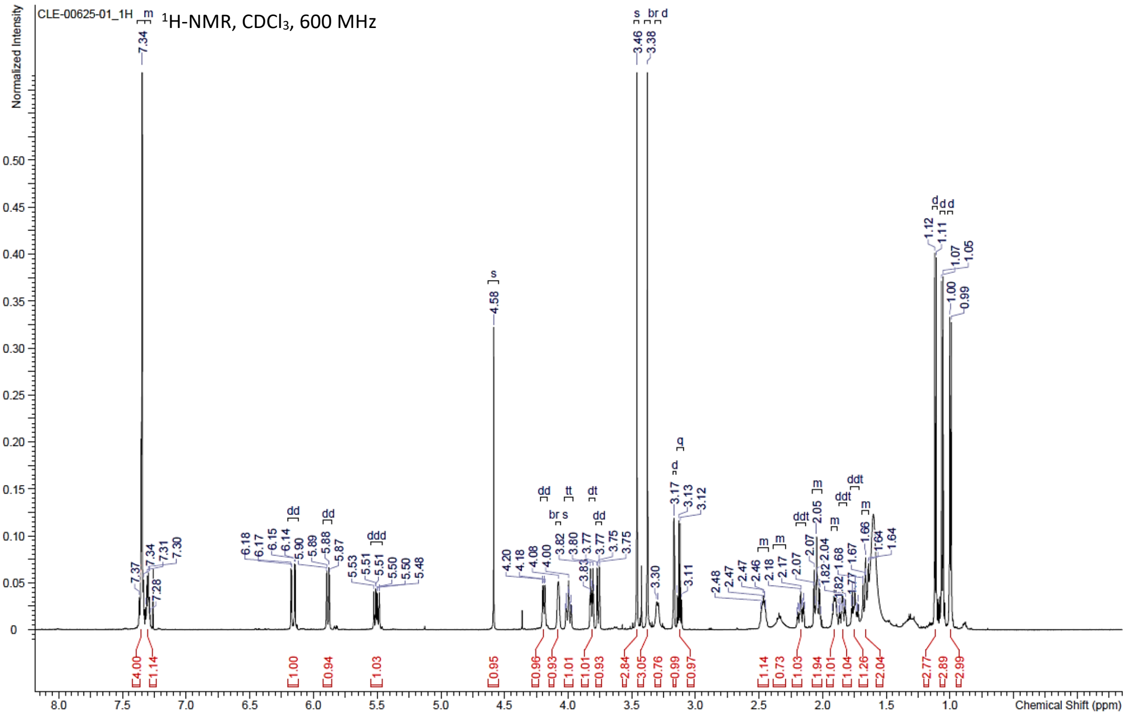

4.16
1.16
1.00
1.00
1.10
1.04
0.95
1.01
0.99
0.99
0.82
3.36
2.99
2.96
1.15
1.79
1.07
0.98
1.04
1.04
2.03
1.46
1.85
3.02
3.21
3.35
F2 Chemical Shift (ppm)
6.0
5.5
5.0
4.5
4.0
3.5
3.0
2.5
2.0
1.5
1.0
8
16
24
32
40
48
56
64
72
80
88
96
104
112
120
128
136
144
152
160
168
176

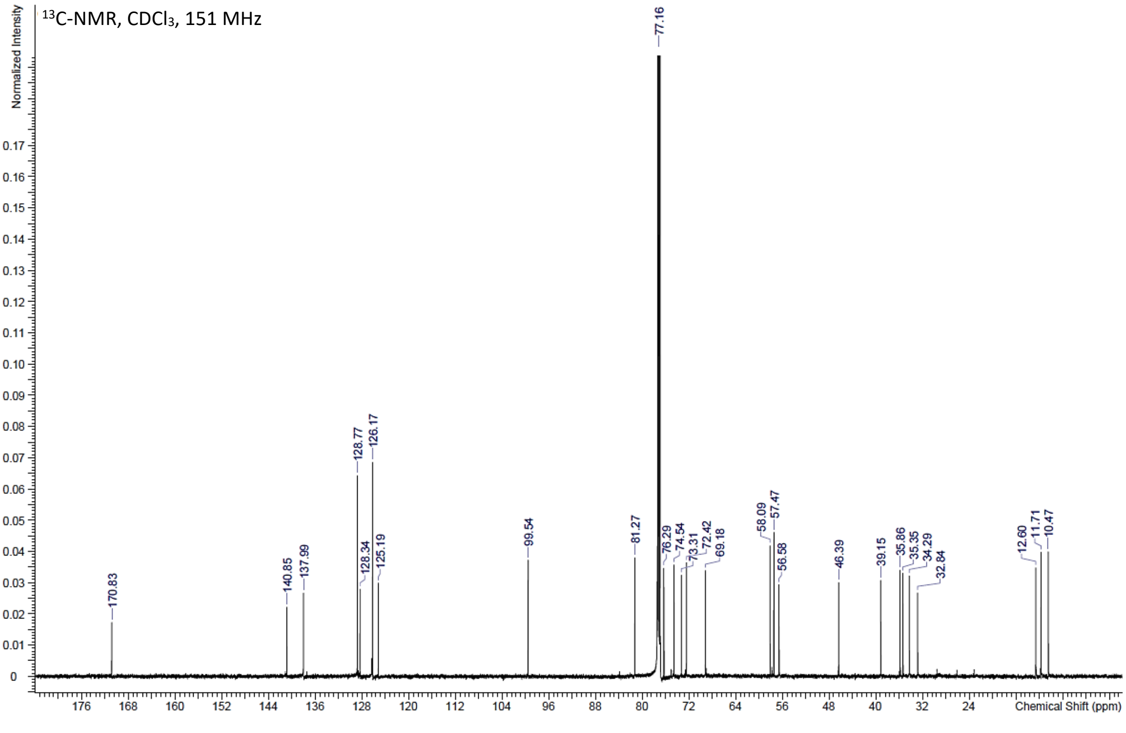
ROESY


4.16
1.16
1.00
1.00
1.10
1.04
0.95
1.01
0.99
0.99
0.82
3.36
2.99
2.96
1.15
1.79
1.07
0.98
1.04
1.04
2.03
1.46
1.85
3.02
3.21
3.35
F2 Chemical Shift (ppm)
6
5
4
3
2
1
0.5
1.0
1.5
2.0
2.5
3.0
3.5
4.0
4.5
5.0
5.5
6.0
6.5
7.0
7.5
8.0

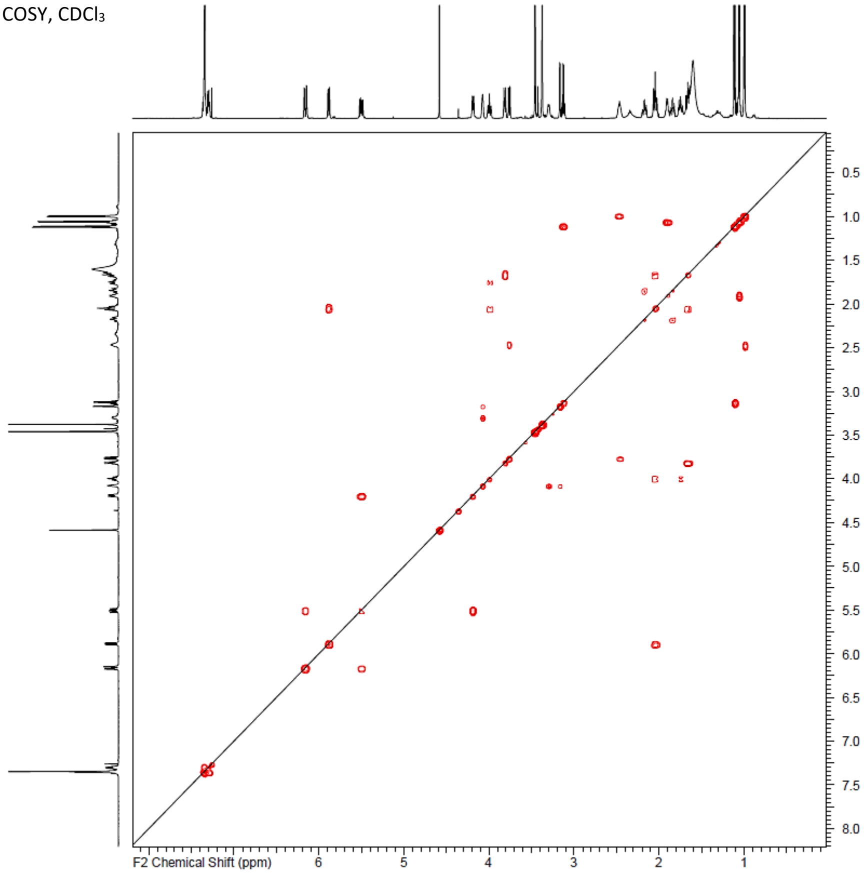
Supplementary Figure 30. NMR of compound **2a**.



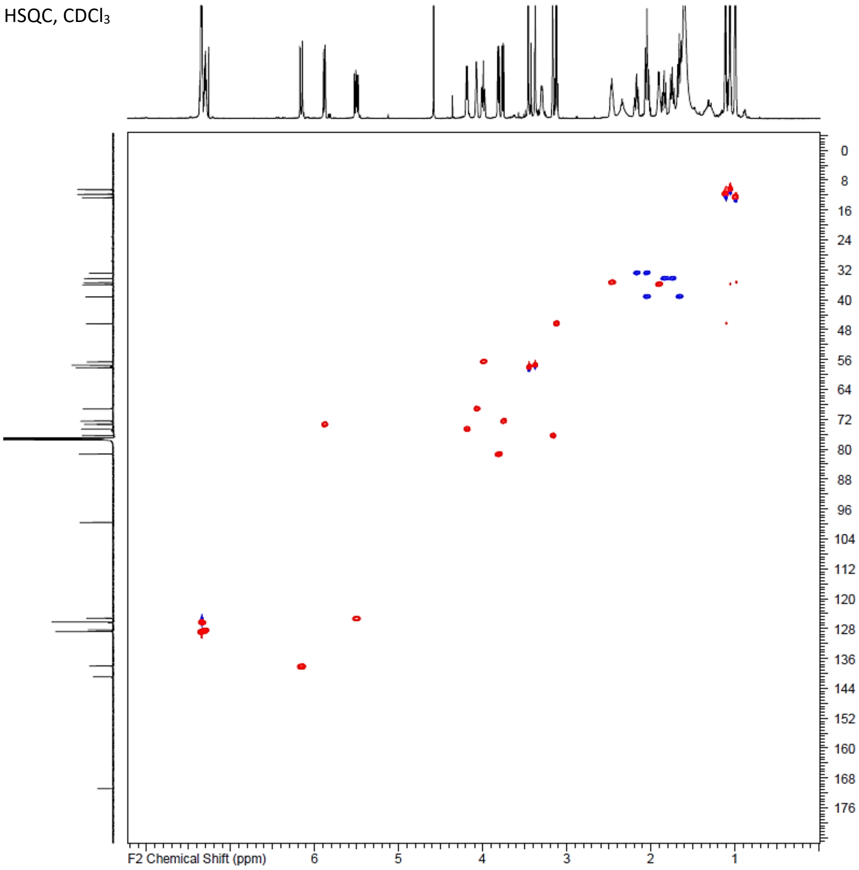




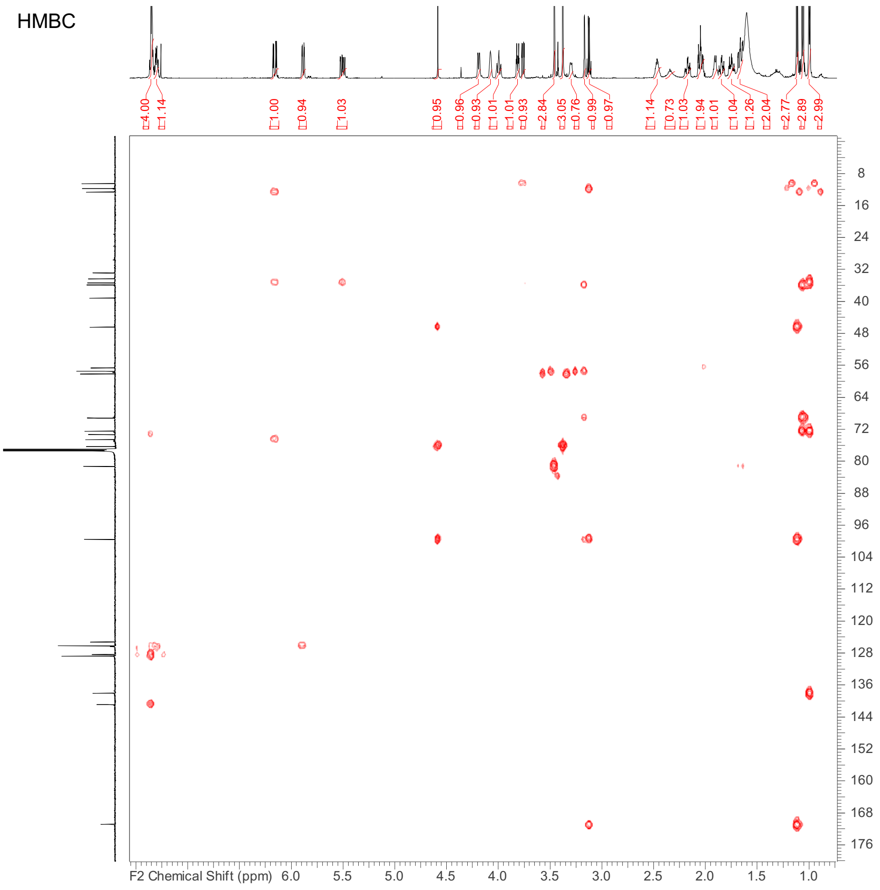
COSY, CDCl₃



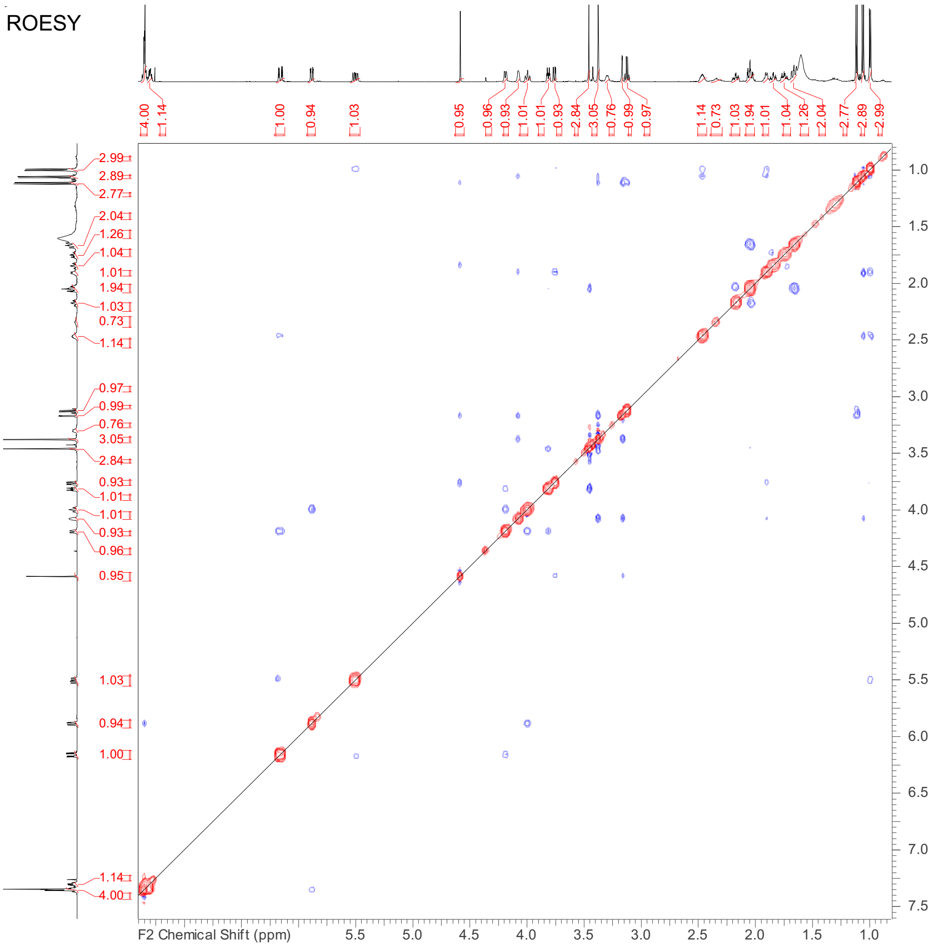
HSQC, CDCl₃



HMBC


4.00
1.14
1.00
0.94
1.03
0.95
0.96
0.93
1.01
1.01
0.93
2.84
3.05
0.76
0.99
0.97
1.14
0.73
1.03
1.94
1.01
1.04
1.26
2.04
2.77
2.89
2.99
8
16
24
32
40
48
56
64
72
80
88
96
104
112
120
128
136
144
152
160
168
176
F2 Chemical Shift (ppm)
6.0
5.5
5.0
4.5
4.0
3.5
3.0
2.5
2.0
1.5
1.0

ROESY


4.00
1.14
1.00
0.94
1.03
0.95
0.96
0.93
1.01
1.01
0.93
2.84
3.05
0.76
0.99
0.97
1.14
0.73
1.03
1.94
1.01
1.04
1.26
2.04
2.77
2.89
2.99
1.0
1.5
2.0
2.5
3.0
3.5
4.0
4.5
5.0
5.5
6.0
6.5
7.0
7.5
F2 Chemical Shift (ppm)
5.5
5.0
4.5
4.0
3.5
3.0
2.5
2.0
1.5
1.0

Methods for biological testing and BP80 determination

The assays were performed at Syngenta's high-throughput screening facilities, using standardized assays and the necessary standards.

Leaf disk assays:

Erysiphe graminis f.sp. tritici (Wheat powdery mildew): preventive application
Barley leaf segments were placed on agar in multiwell plates (24-well format) and sprayed with test solutions. After drying, the leaf disks were inoculated with spores of the fungus. After appropriate incubation the activity of a compound was assessed 7 dpi (days post inoculation) as preventive fungicidal activity.

Puccinia recondita (Brown rust): curative application
Wheat leaf segments are placed on agar in multiwell plates (24-well format). The leaf disks are then inoculated with a spore suspension of the fungus. One day after inoculation the test solution is applied. After appropriate incubation the activity of a compound is assessed 8 dpi (days post inoculation) as curative fungicidal activity.

Liquid culture assays:

Botrytis cinerea (Gray mould):
Conidia of the fungus from cryogenic storage were directly mixed into nutrient broth (Vogel's minimal media). A DMSO solution of the test compounds was placed into a microtiter plate (96-well format) and the nutrient broth containing the fungal spores was added to it. The test plates were incubated at 24 °C and the inhibition of growth was determined photometrically after 72 hours at 620 nm.

Mycosphaerella arachidis (Brown leaf spot of peanut):
Conidia of the fungus from cryogenic storage were directly mixed into nutrient broth (PDB potato dextrose broth). A DMSO solution of the test compounds was placed into a microtiter plate (96-well format) and the nutrient broth containing the fungal spores was added to it. The test plates were incubated at 24 °C and the inhibition of growth was determined photometrically after approximately 5-6 days at 620 nm.

Septoria tritici (leaf blotch):
Conidia of the fungus from cryogenic storage were directly mixed into nutrient broth (PDB potato dextrose broth). A DMSO solution of the test compounds was placed into a microtiter plate (96-well format) and the nutrient broth containing the fungal spores was added to it. The test plates were incubated at 24 °C and the inhibition of growth was determined photometrically after 72 hours at 620 nm.

Monographella nivalis (snow mould, foot rot of cereals):
Conidia of the fungus from cryogenic storage were directly mixed into nutrient broth (PDB potato dextrose broth). A DMSO solution of the test compounds was placed into a microtiter plate (96-well format) and the nutrient broth containing the fungal spores was added to it. The test plates were incubated at 24 °C and the inhibition of growth was determined photometrically after 72 hours at 620 nm.

BP80 determination:
A dilution series was performed for each fungus/compound combination and the efficiency of the compounds evaluated. BP80 represents the breakpoint below which less than 80% efficiency is observed.

The experiments on living organisms were performed as single or triplicate experiments, against positive and negative standards (both commercial and company internal). The triplicate experiments were technical triplicates, with three samples tested in parallel during the same test session. The pest control percentage is assessed visually by a trained personal and is assigned to be 100%, 90%, 70%, 50%, 20% or 0%. The BP80 was determined as an average over all the results.

Supplementary Table 13. BP80 determination for compound **1**.

Compound 1								
Pathogen tested	**Rate of application in ppm**	**Rate of application in uM**	**% Pest control**				**pBP80**	**BP80 (uM)**
			Test 1	**Test 2**				
				replicate 1	replicate 2	replicate 3		
Erysiphe graminis	22.2	42.6	100	100	100	100	5.8	0.16
	7.39	14.2	100	100	100	100		
	2.46	4.7	100	100	100	100		
	0.819	1.6	100	100	100	100		
Puccinia recondita	22.2	42.6	100	100	100	100	5.8	0.16
	7.39	14.2	100	100	100	100		
	2.46	4.7	100	100	100	100		
	0.819	1.6	100	100	100	100		
Botrytis cinerea	6.67	12.8	100	100	100	100	6.8	0.016
	2.22	4.3	100	100	100	100		
	0.739	1.4	100	100	100	100		
	0.246	0.5	100	100	100	100		
Mycosphaerella arachidis	6.67	12.8	100	100	100	100	6.3	0.050
	2.22	4.3	100	100	100	100		
	0.739	1.4	100	100	100	100		
Septoria tritici	6.67	12.8	100	100	100	100	6.3	0.050
	2.22	4.3	100	100	100	100		
	0.739	1.4	100	100	100	100		
	0.246	0.5	100	70	70	70		
Monographella nivalis	6.67	12.8	100	100	100	100	6.8	0.016
	2.22	4.3	100	100	100	100		
	0.739	1.4	100	100	100	100		
	0.246	0.5	100	100	100	100		

Supplementary Table 14. BP80 determination for compound **1a**.

Compound 1a								
Pathogen tested	**Rate of application in ppm**	**Rate of application in uM**	**% Pest control**				**pBP80**	**BP80 (uM)**
			Test 1	**Test 2**				
				replicate 1	replicate 2	replicate 3		
Erysiphe graminis	22.2	42.6	100	90	100	100	4.4	3.98
	7.39	14.2	50	20	50	70		
	2.46	4.7	20	0	0	20		
	0.819	1.6	0	0	0	0		
Puccinia recondita	22.2	42.6	90	20	70	100	4.4	0.40
	7.39	14.2	20	20	0	0		
	2.46	4.7	20	0	0	0		
	0.819	1.6	0	0	0	0		
Botrytis cinerea	6.67	12.8	100	90	100	100	4.9	1.26
	2.22	4.3	50	70	70	70		
	0.739	1.4	20	20	20	20		
	0.246	0.5	20	0	0	0		
Mycosphaerella arachidis	6.67	12.8	90	100	70	90	4.9	1.26
	2.22	4.3	0	20	20	20		
	0.739	1.4	0	0	0	0		
Septoria tritici	6.67	12.8	20	20	70	50	0	1000
	2.22	4.3	0	0	0	0		
	0.739	1.4	0	0	0	0		
	0.246	0.5	0	0	0	0		
Monographella nivalis	6.67	12.8	100	100	90	90	4.9	1.26
	2.22	4.3	50	50	50	50		
	0.739	1.4	20	0	0	0		
	0.246	0.5	20	0	0	0		

Supplementary Table 15. BP80 determination for compound **1b**.

Compound 1b								
Pathogen tested	Rate of application in ppm	Rate of application in uM	% Pest control				pBP80	BP80 (uM)
			Test 1	Test 2				
				replicate 1	replicate 2	replicate 3		
Erysiphe graminis	22.2	42.6	100	100	100	100	5.8	0.16
	7.39	14.2	100	100	100	100		
	2.46	4.7	100	90	100	90		
	0.819	1.6	100	70	50	70		
Puccinia recondita	22.2	42.6	100	100	100	100	4.9	1.26
	7.39	14.2	50	100	90	100		
	2.46	4.7	0	50	90	20		
	0.819	1.6	0	0	0	20		
Botrytis cinerea	6.67	12.8	100	100	100	100	5.4	0.40
	2.22	4.3	90	90	90	100		
	0.739	1.4	50	90	70	70		
	0.246	0.5	0	50	0	0		
Mycosphaerella arachidis	6.67	12.8	100	100	100	100	4.9	1.26
	2.22	4.3	50	90	50	20		
	0.739	1.4	0	0	0	20		
Septoria tritici	6.67	12.8	100	100	100	100	4.9	1.26
	2.22	4.3	70	20	50	90		
	0.739	1.4	50	0	0	20		
	0.246	0.5	0	0	0	0		
Monographella nivalis	6.67	12.8	100	100	100	100	5.4	0.40
	2.22	4.3	100	100	90	100		
	0.739	1.4	70	90	50	20		
	0.246	0.5	20	0	0	20		

Supplementary Table 16. BP80 determination for compound **1c**.

Compound 1c								
Pathogen tested	**Rate of application in ppm**	**Rate of application in uM**	**% Pest control**				**pBP80**	**BP80 (uM)**
			Test 1	**Test 2**				
				replicate 1	replicate 2	replicate 3		
Erysiphe graminis	22.2	42.6	90	70	70	70	4.4	3.98
	7.39	14.2	0	0	50	20		
	2.46	4.7	0	0	0	20		
	0.819	1.6	0	0	0	0		
Puccinia recondita	22.2	42.6	0	0	0	0	0.05	1000
	7.39	14.2	0	0	0	0		
	2.46	4.7	0	0	0	0		
	0.819	1.6	0	0	0	0		
Botrytis cinerea	6.67	12.8	0	0	0	0	0	1000
	2.22	4.3	0	0	0	0		
	0.739	1.4	0	0	0	0		
	0.246	0.5	0	0	0	0		
Mycosphaerella arachidis	6.67	12.8	0	0	0	0	0	1000
	2.22	4.3	0	0	0	0		
	0.739	1.4	0	0	0	0		
Septoria tritici	6.67	12.8	20	0	0	0	0	1000
	2.22	4.3	0	0	0	0		
	0.739	1.4	0	0	0	0		
	0.246	0.5	0	0	0	0		
Monographella nivalis	6.67	12.8	20	0	0	0	0	1000
	2.22	4.3	20	0	0	0		
	0.739	1.4	0	0	0	0		
	0.246	0.5	0	0	0	0		

Supplementary Table 17. BP80 determination for compound **2**.

Compound 2								
Pathogen tested	**Rate of application in ppm**	**Rate of application in uM**	**% Pest control**				**pBP80**	**BP80 (uM)**
			Test 1	**Test 2**				
				replicate 1	replicate 2	replicate 3		
Erysiphe graminis	22.2	42.6	100	100	100	100	5.8	0.16
	7.39	14.2	100	100	100	100		
	2.46	4.7	100	100	100	100		
	0.819	1.6	90	100	100	100		
Puccinia recondita	22.2	42.6	100	100	100	100	5.8	0.16
	7.39	14.2	100	100	100	100		
	2.46	4.7	100	100	100	100		
	0.819	1.6	100	90	100	90		
Botrytis cinerea	6.67	12.8	100	100	100	100	5.8	0.16
	2.22	4.3	100	100	100	100		
	0.739	1.4	50	90	90	90		
	0.246	0.5	20	20	20	20		
Mycosphaerella arachidis	6.67	12.8	100	100	100	100	5.4	0.4.
	2.22	4.3	90	90	90	90		
	0.739	1.4	50	0	20	50		
Septoria tritici	6.67	12.8	90	100	100	100	5.4	0.40
	2.22	4.3	70	70	70	70		
	0.739	1.4	20	0	0	0		
	0.246	0.5	0	0	0	0		
Monographella nivalis	6.67	12.8	100	100	100	100	5.4	0.40
	2.22	4.3	100	90	90	90		
	0.739	1.4	70	50	50	50		
	0.246	0.5	50	0	20	0		

Supplementary Table 18. BP80 determination for compound **2a**.

Compound 2a								
Pathogen tested	**Rate of application in ppm**	**Rate of application in uM**	**% Pest control**				**pBP80**	**BP80 (uM)**
			Test 1	**Test 2**				
				replicate 1	replicate 2	replicate 3		
Erysiphe graminis	22.2	42.6	100	100	100	100	5.8	1.26
	7.39	14.2	100	50	100	50		
	2.46	4.7	0	0	70	0		
	0.819	1.6	0	0	0	0		
Puccinia recondita	22.2	42.6	100	100	100	100	4.9	1.26
	7.39	14.2	50	90	90	100		
	2.46	4.7	0	0	90	0		
	0.819	1.6	0	0	50	0		
Botrytis cinerea	6.67	12.8	50	90	90	90	4.9	1.26
	2.22	4.3	20	50	50	50		
	0.739	1.4	0	0	0	0		
	0.246	0.5	0	0	0	0		
Mycosphaerella arachidis	6.67	12.8	50	20	20	20	0.05	1000
	2.22	4.3	0	0	0	0		
	0.739	1.4	0	0	0	0		
Septoria tritici	6.67	12.8	20	0	0	20	0	1000
	2.22	4.3	0	0	0	0		
	0.739	1.4	0	0	0	0		
	0.246	0.5	0	0	0	0		
Monographella nivalis	6.67	12.8	50	50	70	70	0.05	1000
	2.22	4.3	20	20	20	20		
	0.739	1.4	0	0	0	0		
	0.246	0.5	0	0	0	0		

Combinatorial library WelO5*

Codon list incorporated in the WelO5* site-saturation library. The most abundant codons of *E. coli* were chosen in library design. By replacing the wild-type codons on positions V81, A88, and I161 with the codons in the table, the DNA sequences of the constructed variants can be obtained.

Supplementary Table 19. Codon list incorporated in the WelO5* site-saturation library.

Amino acid	Codon	Amino acid	Codon
A	GCA	M	ATG
C	TGT	N	AAT
D	GAT	P	CCG
E	GAA	Q	CAG
F	TTT	R	CGT
G	GGT	S	AGC
H	CAT	T	ACC
I	ATC	V	GTT
K	AAA	W	TGG
L	CTG	Y	TAT

DNA sequence of the ordered gene fragments:

>Twist_gene_fragment_of_the_WelO5*_site_saturation_library
CCCGTCACCTTTGGCTTATCAGTGAGATATACATATGTCGAACAACACCATCTCGACCAAACCAGCCTTGCATTT
TCTCGACATCAACGCCACCGAAGTCAAGAAATATCCCACTGCAATTCAGGACATCATTATCAATCGCTCATTCGA
TGGCATGATTATTCGGGGAGTCTTTCCTCGCGATACGATGGAGCAGGTTGCTCGTTGCCTGGAAGAAGGGAATGA
TGGCGGCATGAAATCCATCCTGAACAAGAATGAAGAGTTTGGTACGAAA**GTT**GCCCAGATTTATGGCCAT**GCG**AT
TGTTGGCCAATCTCCGGATCTCAAAGACTATTTTGCTAGTTCTGCCATTTTCCGTCAGGCGTGTCGTACCATGTT
TCAGGGTAGCCCGGACTTTGAGGAACAAGTGGAGAGCATTTTCCACTCGTTATCCGGACTGCCCGTAGAGATTCC
GACGGGTCCTGAAGGGCAAACTTACACCCCGGCAACCATTCGTCTGCTGTTAGAAGGCCGCGAA**ATT**GCCGTACA
TGTGGGCAACGACTTTCTTCTGATGCCGGCTGCAAACCATCTGAAAACGTTGCTGGATCTGTCTGATCAACTGTC
GTACTTTATCCCGTTAACAGTGCCGGAAGCAGGTGGTGAATTGGTGGTGTACAACCTGGAATGGAATCCGCAGGA
AGTGGACAAATCAGCGGATCTTCACAAGTACATCGATGAGGTCGAAAGCAAATTCAAAAGCAATCAGAGTCAGAG
TGTTGCGTATGCGCCTGGTCCAGGTGATATGCTCCTGTTCAATGGCGGTCGCTATTATCACCGCGTCAGCGAAGT
AATCGGGAATTCCCCACGTCGCACAATTGGCGGATTTCTGGCGTTCTCAAAAGAGCGCAACAAAATCTACTATTG
GAGCTAACTCGAGCACCAGTGACATCTGGACGCTAAGACCG

green = flanking sequence (Twist)
yellow = flanking sequence including restriction sites
bold = mutation site
underlined = WelO5* gene sequence (ORF)

>WelO5* amino acid sequence
MSNNTISTKPALHFLDINATEVKKYPTAIQDIIINRSFDGMIIRGVFPRDTMEQVARCLEEGNDGGMKSILNKNE
EFGTK**V**AQIYGH**A**IVGQSPDLKDYFASSAIFRQACRTMFQGSPDFEEQVESIFHSLSGLPVEIPTGPEGQTYTPA
TIRLLLEGRE**I**AVHVGNDFLLMPAANHLKTLLDLSDQLSYFIPLTVPEAGGELVVYNLEWNPQEVDKSADLHKYI
DEVESKFKSNQSQSVAYAPGPGDMLLFNGGRYYHRVSEVIGNSPRRTIGGFLAFSKERNKIYYWS-
bold = mutation site

IV. Supplementary References

1. Agarwal, V. *et al.* Biosynthesis of polybrominated aromatic organic compounds by marine bacteria. *Nat. Chem. Biol.* **10**, 640–647 (2014).

2. Zehner, S. *et al.* A regioselective tryptophan 5-halogenase is involved in pyrroindomycin biosynthesis in *Streptomyces rugosporus* LL-42D005. *Chem. Biol.* **12**, 445–452 (2005).

3. Heemstra, J. R. & Walsh, C. T. Tandem action of the O_2- and $FADH_2$-dependent halogenases KtzQ and KtzR produce 6,7-dichlorotryptophan for kutzneride assembly. *J. Am. Chem. Soc.* **130**, 14024–14025 (2008).

4. Menon, B. R. K. *et al.* Structure and biocatalytic scope of thermophilic flavin-dependent halogenase and flavin reductase enzymes. *Org. Biomol. Chem.* **14**, 9354–9361 (2016).

5. Zeng, J. & Zhan, J. Characterization of a tryptophan 6-halogenase from *Streptomyces toxytricini*. *Biotechnol. Lett.* **33**, 1607–1613 (2011).

6. Kirner, S. *et al.* Functions encoded by pyrrolnitrin biosynthetic genes from *Pseudomonas fluorescens*. *J. Bacteriol.* **180**, 1939–1943 (1998).

7. Yeh, E., Garneau, S. & Walsh, C. T. Robust *in vitro* activity of RebF and RebH, a two-component reductase/halogenase, generating 7-chlorotryptophan during rebeccamycin biosynthesis. *Proc. Natl. Acad. Sci. U. S. A.* **102**, 3960–3965 (2005).

8. Seibold, C. *et al.* A flavin-dependent tryptophan 6-halogenase and its use in modification of pyrrolnitrin biosynthesis. *Biocatal. Biotransformation* **24**, 401–408 (2006).

9. Wang, S. *et al.* Functional characterization of the biosynthesis of Radicicol, an Hsp90 inhibitor resorcylic acid lactone from Chaetomium chiversii. *Chem. Biol.* **15**, 1328–1338 (2008).

10. Fraley, A. E. *et al.* Function and structure of MalA/MalA', iterative halogenases for late-stage C-H functionalization of indole alkaloids. *J. Am. Chem. Soc.* **139**, 12060–12068 (2017).

11. Zeng, J. & Zhan, J. A novel fungal flavin-dependent halogenase for natural product biosynthesis. *ChemBioChem* **11**, 2119–2123 (2010).

12. Neumann, C. S., Walsh, C. T. & Kay, R. R. A flavin-dependent halogenase catalyzes the chlorination step in the biosynthesis of *Dictyostelium* differentiation-inducing factor 1. *Proc. Natl. Acad. Sci. U. S. A.* **107**, 5798–5803 (2010).

13. Shepherd, S. A. *et al.* Extending the biocatalytic scope of regiocomplementary flavin-dependent halogenase enzymes. *Chem. Sci.* **6**, 3454–3460 (2015).

14. Payne, J. T., Poor, C. B. & Lewis, J. C. Directed evolution of RebH for site-selective halogenation of large biologically active molecules. *Angew. Chem. Int. Ed.* **54**, 4226–4230 (2015).

15. Andorfer, M. C., Park, H. J., Vergara-Coll, J. & Lewis, J. C. Directed evolution of RebH for catalyst-controlled halogenation of indole C-H bonds. *Chem. Sci.* **7**, 3720–3729 (2016).

16. Poor, C. B., Andorfer, M. C. & Lewis, J. C. Improving the stability and catalyst lifetime of the halogenase RebH by directed evolution. *ChemBioChem* **15**, 1286–1289 (2014).

17. Hillwig, M. L. & Liu, X. A new family of iron-dependent halogenases acts on freestanding substrates. *Nat. Chem. Biol.* **10**, 921–923 (2014).

18. Zhu, Q. & Liu, X. Characterization of non-heme iron aliphatic halogenase WelO5* from Hapalosiphon welwitschii IC-52-3: identification of a minimal protein sequence motif that confers enzymatic chlorination specificity in the biosynthesis of welwitindolelinones. *Beilstein J. Org. Chem.* **13**, 1168–1173 (2017).

19. Hillwig, M. L., Zhu, Q., Ittiamornkul, K. & Liu, X. Discovery of a promiscuous non-heme iron halogenase in ambiguine alkaloid biogenesis: implication for an evolvable enzyme family for late-stage halogenation of aliphatic carbons in small molecules. *Angew. Chem. Int. Ed.* **55**, 5780–5784 (2016).

20. Mitchell, A. J. *et al.* Structure-guided reprogramming of a hydroxylase to halogenate its small molecule substrate. *Biochemistry* **56**, 441–444 (2017).

21. Fox, R. J. *et al.* Improving catalytic function by ProSAR-driven enzyme evolution. *Nat. Biotechnol.* **25**, 338–344 (2007).

22. Feng, X., Sanchis, J., Reetz, M. T. & Rabitz, H. Enhancing the efficiency of directed evolution in focused enzyme libraries by the adaptive substituent reordering algorithm. *Chem. Eur. J.* **18**, 5646–5654 (2012).

23. Saito, Y. *et al.* Machine-learning-guided mutagenesis for directed evolution of fluorescent proteins. *ACS Synth. Biol.* **7**, 2014–2022 (2018).

24. Cadet, F. *et al.* A machine learning approach for reliable prediction of amino acid interactions and its application in the directed evolution of enantioselective enzymes. *Sci. Rep.* **8**, (2018).

25. Wu, Z., Jennifer Kan, S. B., Lewis, R. D., Wittmann, B. J. & Arnold, F. H. Erratum: Machine learning-assisted directed protein evolution with combinatorial libraries. *Proc. Natl. Acad. Sci. U. S. A.* **117**, 788–789 (2020).

26. Vornholt, T. *et al.* Systematic engineering of artificial metalloenzymes for new-to-nature reactions. *Sci. Adv.* **7**, eabe4208 (2021).

27. Waterhouse, A. *et al.* SWISS-MODEL: homology modelling of protein structures and complexes. *Nucleic Acids Res.* **46**, W296–W303 (2018).

28. Jumper, J. *et al.* Highly accurate protein structure prediction. *Nature* (2021). doi:10.1038/s41586-021-03819-2

29. Trott, O. & Olson, A. J. AutoDock Vina: Improving the speed and accuracy of docking with a new scoring function, efficient optimization, and multithreading. *J. Comput. Chem.* **31**, 455–461 (2010).

30. Stourac, J. *et al.* Caver Web 1.0: Identification of tunnels and channels in proteins and analysis of ligand transport. *Nucleic Acids Res.* **47**, W414–W422 (2019).

31. Zirkle, R., Ligon, J. M. & Molnár, I. Heterologous production of the antifugal polyketide antibiotic soraphen A of *Sorangium cellulosum* So ce26 in Streptomyces lividans. *Microbiology* **150**, 2761–2774 (2004).

32. Bedorf, N. *et al.* Mikrobiologisches Verfahren zur Herstellung agrarchemisch verwendbarer mikrobizider makrozyklischer Lactonderivate. EP 358606 A2 (1990).

33. Sutter, M., O'Sullivan, A., Hoefle, G., Boehlendorf, B. & Kiffe, M. Sopharen c Markolid-Derivate und ihre Verwendung als Mikrobizide. EP540469 (1993).

Article III

LibGENiE – A bioinformatic pipeline for the design of information-enriched enzyme libraries

David Patsch[a,b], Michael Eichenberger[a], Moritz Voss[a], Uwe T. Bornscheuer[b] and Rebecca M. Buller[a,*]

[a] Zurich University of Applied Sciences, School of Life Sciences and Facility Management, Institute of Chemistry and Biotechnology, Einsiedlerstrasse 31, 8820 Wädenswil, Switzerland

[b] Institute of Biochemistry, Department of Biotechnology & Enzyme Catalysis, Greifswald University, Felix-Hausdorff-Strasse 4, D17487 Greifswald, Germany

* Corresponding author: Rebecca M. Buller (rebecca.buller@zhaw.ch)

Abstract

Enzymes are potent catalysts with high specificity and selectivity. To leverage Nature's synthetic potential for industrial applications, various protein engineering techniques have emerged which allow to tailor the catalytic, biophysical, and molecular recognition properties of enzymes. However, the many possible ways a protein can be altered forces researchers to carefully balance between the exhaustiveness of an enzyme screening campaign and the required resources. Consequently, the optimal engineering strategy is often defined on a case-by-case basis. Strikingly, while predicting mutations that lead to an improved target function is challenging, here we show that the prediction and exclusion of deleterious mutations is a much more straightforward task as analyzed for an engineered carbonic acid anhydrase, a transaminase, a squalene-hopene cyclase and a Kemp eliminase. Combining such a pre-selection of allowed residues with advanced gene synthesis methods opens a path toward an efficient and generalizable library construction approach for protein engineering. To give researchers easy access to this methodology, we provide the website LibGENiE containing the bioinformatic tools for the library design workflow.

Keywords: Bioinformatic Tools; Enzyme Engineering**;** Library Design; Sequence Space

1. Introduction

Enzymes are remarkable catalysts capable of facilitating complex reactions with high substrate specificity and exquisite chemo-, regio- and enantioselectivity [1]. However, when used in conditions necessary to drive a process at an industrial scale, the performance of wild-type enzymes often remains insufficient from an economic standpoint. Thus, to better harness the capabilities of Nature's catalysts in industrial settings, much focus has been placed on advancing protein engineering strategies to proficiently tailor enzymes' catalytic, biophysical, and molecular recognition properties [2,3]. In this way, enzyme engineering has allowed to broaden the substrate scope of natural enzymes [4], change their chemistry [5], improve catalytic activity [6–8], or alter enantioselectivity [9,10]. Yet, despite their successful outcome, these protein engineering examples did not explore all possible amino acid configurations of the target enzymes, and consequently, the solutions found in evolution campaigns might be far from optimal. However, since the number of possible enzyme variants scales exponentially with protein sequence length, the screening burden imposed on researchers quickly becomes intractable when attempting to explore enzyme composition comprehensively. For illustration, a protein composed of only 100 amino acids can be altered in 20^{100} ways, an astronomical number far exceeding even the estimated number of atoms in the universe [11]. Faced with this challenge, also called "the numbers problem in directed evolution" [12], protein engineers aim to navigate sequence space as efficiently as possible and constantly seek to develop novel methods to optimize the process. Existing approaches can broadly be classified into the categories of 1) directed evolution, 2) semi-rational, and 3) rational protein design (Figure 1) and are often employed in accordance with the available screening capabilities and prior information about the enzymatic system [13].

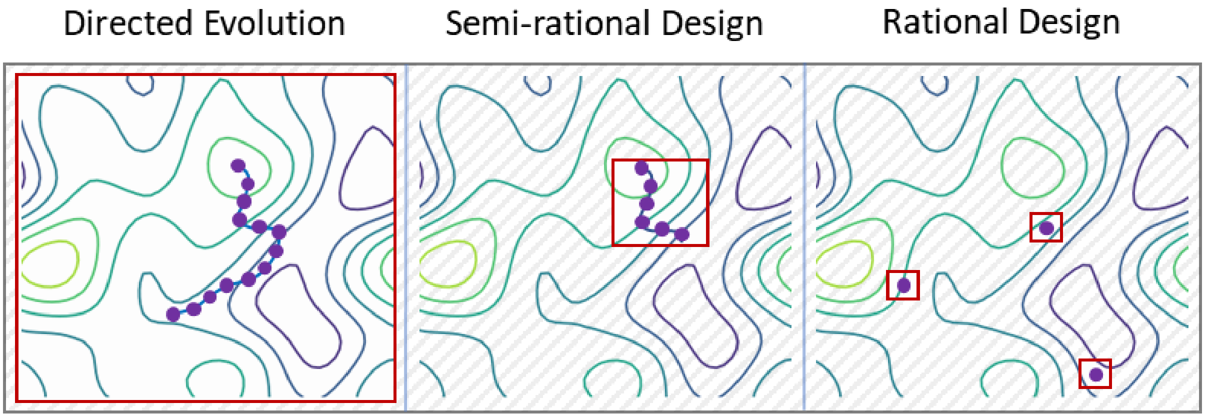


Figure 1: Overview of protein engineering techniques. The different categories are sorted by their required screening effort from left (highest) to right (lowest). In traditional directed evolution, the sequence space (red box) is commonly explored randomly, with little additional information required. Rational design can be viewed as a complementary approach. Information about the system, which can include experimental data, knowledge of the mechanism, as well as computational techniques, is used to reduce the sequence space as much as possible, and areas within it are sampled selectively. Semi-rational design also relies on additional information to reduce the screening space; however, experiments and physical evaluation are still required. Notably, the boundaries between these techniques are often fluid, and the optimal engineering method depends on many factors, such as the complexity of the functional assay, available screening capabilities, or previous knowledge of the enzyme. Image inspired by Bornscheuer *et al.* [13].

Traditional directed evolution, which relies on gene recombination or whole-gene error-prone PCR to create diversity, is often associated with a heavy screening burden [14] as many of the introduced mutations in the libraries are either neutral or unfavorable [15]. Positively, however, directed evolution does not require any prior knowledge about enzyme function or structure to be effective. In contrast, rational enzyme design [16] aims to limit enzymatic screening efforts to only a few distinct amino acid substitutions [17]. The approach relies on an intimate knowledge of a protein's function and/or structure and, as such, requires high predictive accuracy, which can be obtained – at least in part – through the interpretation of experimental data. Although bioinformatic tools such as AlphaFold 2 [18] have facilitated the access to high quality protein models, rational modulation of crucial residues often requires far more fine-grained information on receptor-ligand interaction networks and dynamics. Additionally, significant *in-silico* efforts might be required to resolve uncertainty around specific mechanisms and illuminate required factors between interaction partners to drive a desired reaction [19]. Even with the advanced bioinformatic methods available today, it can be challenging to rationalize which sites, specific residues, or combinations should be selected when optimizing a protein for a certain task.

Lastly, semi-rational protein engineering fuses elements of rational design and directed evolution to create more focused enzyme libraries of higher quality [20,21]. This combination leads to a more efficient sampling of the sequence space, resulting in a lower screening burden than completely random approaches [22,23] while allowing more leniency for computational limitations and inaccuracies. For example, researchers can investigate the 3D structure of an enzyme to identify the catalytic pocket and focus their engineering efforts only on this region which is likely to react more directly to amino acid exchanges. In this way, sequence space can be reduced while beneficial mutations can be largely sampled, as many of them are typically situated in the active site [24]. In practice, researchers often aggregate information from sources such as the target enzyme's 3D structure, function, previous knowledge (for example, mutational data), phylogeny, docking, or machine learning to preselect potential hotspots [16,20]. Based on this information, focused libraries ranging in size from ~200-2000 enzyme variants are constructed. Such screening efforts are within the scope of what GC or HPLC systems can handle within a reasonable timeframe [25]. It should be noted, though, that semi-rational enzyme design also suffers from the "numbers problem in directed evolution", and in many cases, only a small fraction of the targeted variants can be analyzed experimentally. In addition, experimental throughput is hampered by limitations in the physical construction of complex gene libraries.

Using standard molecular biology strategies, the creation of large, randomized libraries through methods such as error-prone PCR or the construction of a few specific variants through site-directed mutagenesis is easily possible. However, building large libraries made up of predefined enzyme variants often remains expensive and challenging. One exciting prospect to address the existing library construction bottlenecks is the use of micro-array-based "oligo-pools". These pools are mixes of up to several hundred thousand individually designed polynucleotides with <300 bp length, synthesized through phosphoramidite chemistry [26]. Notably, array-based oligo synthesis is orders of magnitude cheaper than traditional column-based synthesis routes, with costs ranging from US$ 0.00001–0.001 per nucleotide, depending on length, scale, platform, or vendor [27]. Considering a typical library size for semi-rational enzyme design (< 2000 variants) and a protein of approximately 300 amino acid length, oligo pools for focused libraries can consequently be ordered for roughly 2000 US$ [28], leading to material costs of approximately 1 US$ per variant. Consequently, despite issues like truncated DNA molecules and high error rates [29], the oligo-pool option could be more cost-effective than degenerate or reduced codon coverage primers traditionally employed for library construction strategies while allowing for much more flexibility in library design.

Relevant enzymatic properties to be optimized for industrial applications include activity, thermo- and solvent stability, selectivity, and specificity [30]. As delineated above, reliably selecting appropriate amino acid residues for randomization to improve any of these traits is a challenging aspect of semi-rational enzyme design. Guiding principles might be to select residues near the binding pocket to engineer enantioselectivity [31] or substitute specific residues to redesign unstable protein regions to improve thermostability [32]. Especially the latter, namely the modulation of protein stability through the introduction of mutations, is a widely pursued goal, and different computational procedures have been established to this end, including the use of sophisticated physical force fields, deep learning, and hybrid approaches [33–38]. Nevertheless, today, these tools' performance to accurately predict stabilizing mutations is often unsatisfactory [39].

Intriguingly, computational techniques can be helpful in ways that might not be immediately obvious. For example, we followed the logic that it seems much easier to predict destabilizing mutations than amino acid changes that stabilize a protein scaffold [40]. We consequently reasoned that methods developed to predict enzyme sequences with improved stability might be used in a much broader sense if they were uniquely used to identify destabilizing mutations. Through the exclusion of such destabilizing mutations, the design of solution-enriched enzyme libraries for the optimization of enzyme activity or any other desirable traits would be made possible. The resulting complex libraries could then, in turn, effectively be built using specifically designed oligo-pools.

2. Results

2.1. Predicting (and excluding) destabilizing mutations

To set the basis for our approach, we analyzed available literature data of successful evolution campaigns, including data generated during the optimization of a carbonic anhydrase [7], a transaminase [41], a squalene-hopene cyclase [42] and a Kemp eliminase [6]. In a first step, we calculated the ΔΔG values, a measure of free energy changes upon mutation [43], for all possible amino acid substitutions at all sites in the selected wild-type enzymes using a cartesian ΔΔG protocol implemented in the Rosetta Protein Modelling Suite [44]. For example, in the case of an enzyme consisting of 300 amino acids, all possible 20 * 300 ΔΔG values were calculated. These ΔΔG values can help approximate how mutations affect protein stability by comparing the free energy of the native and altered conformation of a protein. Negative values typically refer to a stabilizing mutation, while strongly positive values denote destabilizing mutations.

Following this protein-wide stability profiling, we analyzed in which range the ΔΔG values of the experimentally determined beneficial mutations of the selected enzymes were located: For example, we studied data generated by Codexis, a US-based company specialized in protein engineering, which evolved a carbonic anhydrase towards improved activity at higher temperatures. To do so, the researchers saturated all non-catalytic residues in a first evolution round [7], identifying 84 unique carbonic anhydrase variants that performed better than the wild-type under their screening conditions. Our ΔΔG analysis indicated that most of the mutations observed in improved variants were within the lowest (stabilizing) 60 % of predicted ΔΔG values hinting that a large part of the screening space could have been excluded *a priori* (Figure 2b). Interestingly, we noted that while we could identify destabilizing mutations, the predicted ΔΔG values became much less informative beyond a certain exclusion threshold. In the ΔΔG range where most improved enzyme variants were found (-7.5 to 4.7 Rosetta energy units (REU), Figure 2a), the measured fold improvement over wild-type did not show a correlation to the calculated ΔΔG values (Pearson correlation coefficient 0.006, Figure 2a).

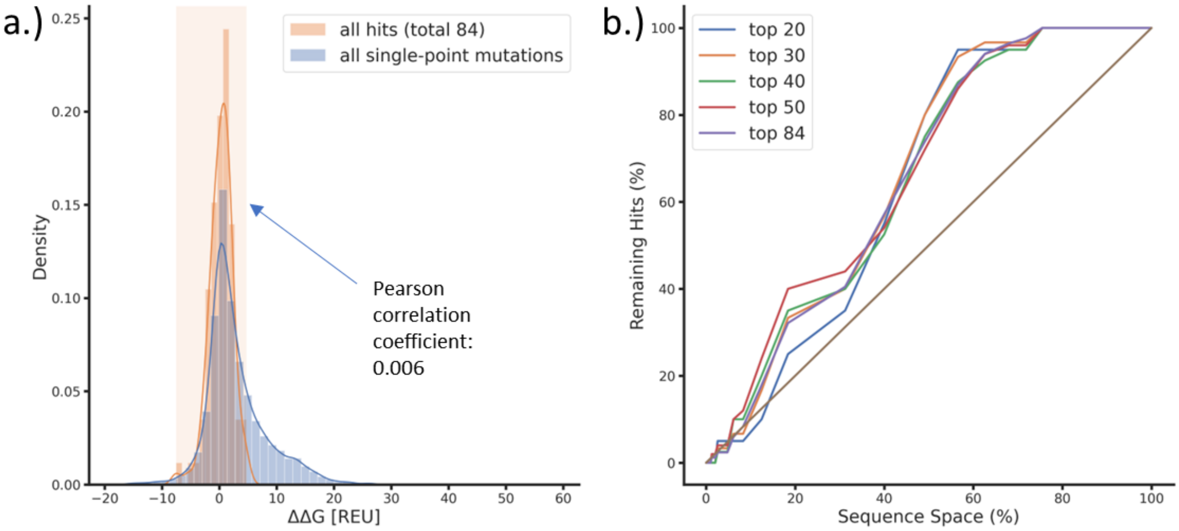


Figure 2: a.) Density plot of predicted ΔΔG values (lower values correspond to higher predicted stability) of a carbonic anhydrase [7]. The blue density curve depicts the ΔΔG values of all possible single-point mutants, and the orange plot represents the ΔΔG distribution of the 84 reported hits. The ΔΔG range in which hits were identified is highlighted in orange. Additionally, the Pearson correlation coefficient between the activity of identified hits and predicted ΔΔG is shown. b.) Line chart of the same dataset as in a.). The x-axis refers to the sequence space when reducing it only through predicted ΔΔG values. For example, if we remove the variants with the highest 10 % of predicted ΔΔG values (most destabilizing), 90 % of the sequence space remains. The y-axis represents how many of the 84 reported hits can be found in a given remaining sequence space. For example, none of the 84 reported hits are within the sequence space characterized by the highest 10 %

predicted ΔΔG values. This analysis is shown for the 20, 30, 40, 50, and 84 best-measured hits (out of 84). As a comparison, the brown line highlights the impact of reducing the sequence space randomly.

To test the general applicability of this finding with examples from distinct enzyme families beyond enzyme class 4 (carbonic anhydrase), we turned to analyze the evolutionary trajectories of enzymes stemming from enzyme class 2 (transaminase), enzyme class 5 (squalene hopene cyclase) as well as a computationally designed enzyme (Kemp eliminase) based on a scaffold from enzyme class 3 (xylanase). The transaminase ATA-217, engineered towards synthesizing a chiral precursor of sacubitril, an active ingredient in the blockbuster drug Entresto, harbored 26 mutations in the final variant [41] whereas four mutations allowed the squalene hopene cyclase *Aci*SHC to gain enantio-complementary access to valuable monocyclic terpenoids [42]. Kemp eliminase HG3, a computationally designed enzyme capable of catalyzing a proton abstraction reaction from 5-nitrobenzisoxazole, was optimized in 17 rounds of directed evolution to yield a variant with 17 mutations whose catalytic activity rivals that of natural enzymes (k_{cat} = 700 ± 60 s^{-1}, $k_{\text{cat}}/K_{\text{m}}$ = 230,000 ± 20’000 $\text{s}^{-1}\ \text{M}^{-1}$) [6].

In all investigated evolution projects, we observed the general trend that destabilizing mutations were not incorporated in evolved enzyme variants. Notably, when comparing amino acid mutations predicted to be destabilizing as single point mutations in the wild-type enzymes to any reported beneficial single point mutation within the evolution campaigns (Figure S1/S2), we deduced that almost all the destabilizing mutations could be excluded confidently at the outset of the enzyme optimization projects (Table 1, Table S1, Figure S1, Figure S2). Interestingly, in the case of evolved *Aci*SHC, we observed a single outlier: Mutation A169P was flagged as destabilizing (21.5 REU) yet still appeared in the optimized squalene-hopene cyclase variant. Potentially, the destabilizing mutation was incorporated because *Aci*SHC is a thermophilic enzyme whose scaffold would generally allow for more leeway toward introducing destabilizing mutations.

Conclusively, the relationship between activity and stability is often complex, with reports of both negative [45–48] and positive correlations [49,50] between stability and function attesting to the fact that different enzymatic systems behave differently to mutations. Strikingly, as highlighted in this work, employing the opposite approach for the construction of information-enriched libraries seems much more reliable: Strongly destabilizing mutations are often accompanied by a loss in function (Table 1, Figure 2), consequently enabling their early exclusion from the sequence pool.

Sequence space (%)												
Dataset		**# Mut**	**100**	**90**	**80**	**70**	**60**	**50**	**40**	**30**	**20**	**10**
	ATA217	26	100	100	96.2	92.3	88.5	73.1	73.1	61.5	53.8	42.3
	HG3.17	17	100	100	100	100	100	82.4	52.9	52.9	47.1	41.2
	DvCA	36	100	100	100	97.2	91.7	77.8	61.1	44.4	38.9	13.9
	AciSHC	4	100	75	75	75	75	75	75	50	50	50
	average		100	93.8	92.8	91.1	88.8	77.1	65.5	52.2	47.4	36.8

Table 1: Overview of how ΔΔG values of single mutations found in the final improved variants of the selected evolution campaigns are distributed within the context of all possible calculated ΔΔG values for the wild-type enzymes. In this analysis, the most destabilizing mutations in the context of the wild-type enzyme are gradually removed (in 10 % steps), reducing the theoretical sequence space from left to right. The remaining sequence space is analyzed with respect to its harboring the amino acid substitutions found in evolved enzyme variants and the value is given in percent (%). For example, in the case of HG3 evolution, a focused library in which the 40 % most destabilizing mutations are removed from sequence space would still contain all the 17 beneficial mutations identified in the final variant.

2.2. Oligo pools for library creation

Promisingly, as seen above, reducing the amino acid alphabet in gene library preparation can be facilitated through computational techniques. Yet, it is equally important to have in mind that such a process might lead to libraries that are too diversified to be easily and economically constructed. In this respect, it is important to consider the redundant nature of the genetic code in which the 20 natural amino acids are encoded by 61 sense codons. In consequence, researchers have tried to avoid using the heavily redundant NNN codon in library construction which additionally suffers from the occurrence of stop codons (N standing for any of the four DNA bases). Instead, they have turned to using primers harboring degenerate codons such as NNK (32 codons, 20 amino acids), NDT (12 codons, 12 amino acids) or using the 22c (22 codons, 20 amino acids), and 20c (20 codons, 20 amino acids) tricks [12,51,52].

Unfortunately, the current strategies using degenerate codons are not suitable to build the information-enriched libraries stemming from our computational workflow, in which each targeted mutation site would demand the inclusion of only certain amino acids (Figure S3). Thus, we set out to evaluate the feasibility of using micro-array-synthesized oligonucleotides, commercially available under the term "oligo-pools", for constructing the complex libraries derived from our stability filtering strategy (Figure 3a, Figure S3).

In particular, we opted to focus our attention on single-point residue exchanges. As there are limitations to the synthesis length of oligo-pools [29], desired mutations must be split across multiple fragments or "sub-pools" (Figure 3a), which can be separated from the main pool with sub-pool specific primers. These sub-pools consist of individual oligonucleotides, each carrying a single mutation, which can be introduced into the gene of interest through traditional molecular biology techniques, such as gene splicing by overlap extension PCR (SOEing) [53].

To evaluate the suitability of the oligo-pools for the construction of tailored enzyme variant libraries, we ordered a pool of 200 oligo sequences encoding the initial 157 bases of the Kemp eliminase HG3 [6]. To create diversity for sequence analysis, three consecutive adenine nucleotides were introduced within four spatially distinct regions of the 157 bp gene fragment (sequence A: bp 30 – 32; sequence; bp 62 – 64; sequence C: bp 93 – 94; sequence D; bp 124- 126) and each such sequence was ordered in the pool fifty times. Following fragment amplification and cloning, we noted relatively high rates (~50 %) of undesirable sequences, split between either wild-type sequences or multiple-point mutants (Figure 3b). This high fraction of incorrect sequences was not wholly unexpected and correlates to the range reported in previous projects that leverage oligo pools for single-point mutation library creation [54–56].

Oligo pools suffer from the low concentration of individual oligonucleotides [29] making an initial amplification step indispensable [57]. In fact, depending on the number of projects combined within one oligo pool, it might be required to perform this amplification twice: once to isolate the sub-pools [58] and then again to separate the individual fragments. We suspected that these PCR amplification steps introduce additional errors into the oligo-pool libraries through uncoupling events that lead to truncated PCR products. These truncated gene products can serve as primers during the next PCR cycle [59,60], either picking-up additional mutations (leading to multiple-point variants) or overwriting desired mutations altogether (resulting in wild-type). As the prevalence of PCR abortions is affected by multiple factors, such as the concentration of nucleotides, the number of PCR cycles, and the polymerase used for amplification [61], we opted to optimize the amplification procedure.

To do so, we investigated ways how to improve the overall sampling efficiency of oligo-libraries by testing different polymerases (Q5, Phusion, and KAPA polymerase), dNTP concentrations, and varying amounts of PCR cycles (15, 30, and 45) for their impact on the formation of undesired gene fragments. Using the same oligo-pool analysis set-up as described previously, it became clear that neither the dNTP concentration nor the number of PCR cycles significantly impacted the number of corrupted sequences (Figure S4). However, the choice of polymerase showed an influence on gene fragment integrity (Figure 3b): While Q5 and Phusion polymerase led to 47.5 – 60 % correct fractions, KAPA polymerase was found to be most suited for oligo-pool amplification (> 60 % correct fragments). The remaining undesirable sequences were split between wild-type (28.3 %) and primarily double-point mutants (8.1 %) (Figure 3c). In summary, we advise that these rates should be considered when designing the sampling strategy of directed evolution studies.

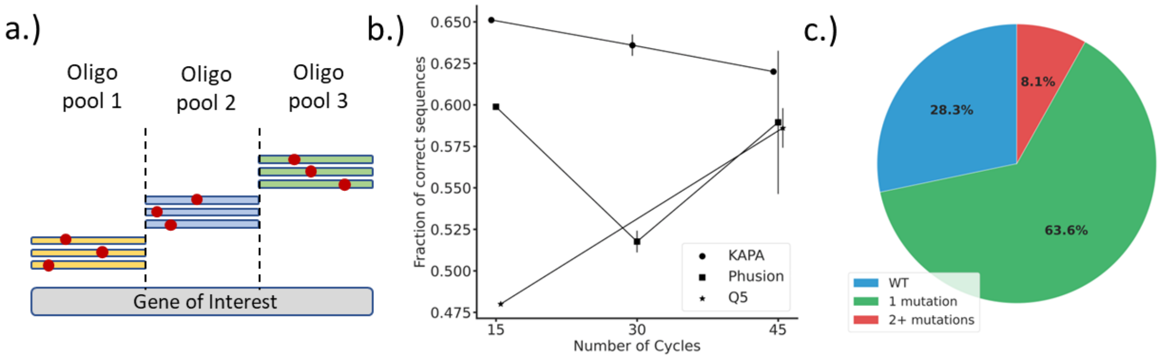


Figure 3: a.) As oligonucleotides ordered within oligo-pools are limited to <300 bp in length, the target gene must be split into smaller fragments below this size. These mutations can then be introduced into the desired gene through standard molecular biology techniques such as SOEing [53]. b.) Fraction of correct sequences in the amplified oligo-pool. The experiments were conducted with varying amounts of PCR cycles (15, 30 and 45), as well as different polymerases (Q5, Phusion, KAPA). The error bars denote the average and error of experiments that vary in their dNTP concentration. c.) Overview of library quality resulting from fragment amplification with KAPA polymerase using 30 amplification cycles. Sequencing highlighted that 63.6 % of variants were produced correctly (one desired mutation – green), while 28.3 % wildtype sequences (blue) and 8.1 % sequences that contain two or more mutations were observed (red).

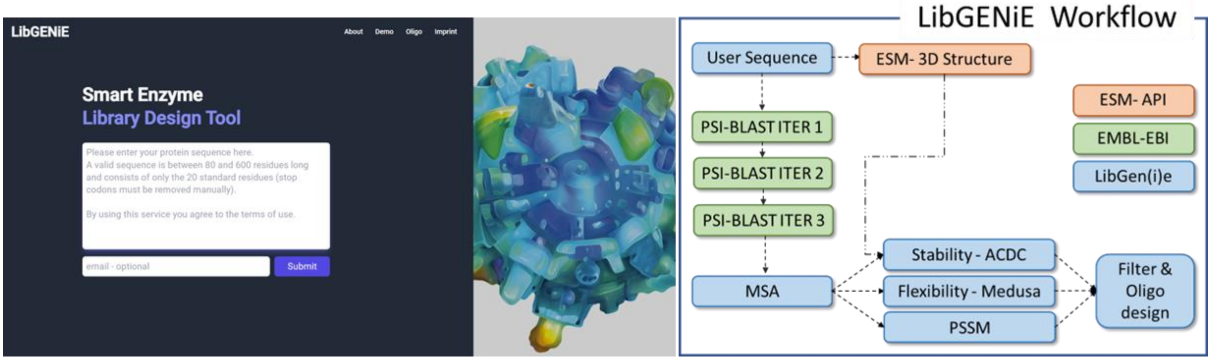
3. LibGENiE: A webserver for smart library creation

To facilitate the design of solution-enriched gene libraries and their subsequent construction with the oligo-pool technique, we set up a web server named LibGENiE (available at www.libgenie.ch).

[the site is password protected during the manuscript evaluation, please log in using the log in credentials:
User: tester@draft.ch
Password: review
this authentication process will be removed upon publication].

LibGENiE provides data sets compiling common protein properties relied upon in rational design, including phylogenetic conservation (extracted from a multi-sequence alignment generated from three rounds of PSI-BLAST with default settings [62]), stability (predicted from protein free energy changes upon point mutations, ACDC-NN [63]), and flexibility (generated from MEDUSA [64]). These tools were chosen based on open access (e.g., license situation) and computational demands. In addition, LibGENiE allows to generate custom oligonucleotides for library construction (Figure 4), which can be designed based on the preceding *in-silico* filtering. In addition, based on a selected maximum gene length, LibGENiE splits the input sequence into even sections and designs the required amplification primers.

Initializing LibGENiE only requires the user to provide a protein sequence. From this, a sequence alignment for the input sequence is generated through three rounds of iterative PSI-BLAST [62]. As detailed below, the multiple sequence alignment then serves as the foundation for all further processing.



Please enter your protein sequence here.
A valid sequence is between 80 and 600 residues long and consists of only the 20 standard residues (stop codons must be removed manually).

By using this service you agree to the terms of use.

Figure *4*: Schematic overview of the LibGENiE landing page and workflow. Based on the user input sequence, three rounds of PSI-BLAST are performed through the EMBL-EBI API [62]. The acquired multiple sequence alignment (MSA) information is then further processed to predict stability (ACDC-NN [63]), flexibility (MEDUSA ([64]), and conservation (MSA from PSI-BLAST). LibGENiE provides raw access to this data, which can be used to restrict the sequence space. In addition, LibGENiE offers a tool for the design of oligo sequences.

3.1.1. Thermodynamic stability

Quantifying the change in free energy between the wild-type protein and a single point variant is mainly associated with expression or stability optimization; however, as delineated above, knowing which residues completely destabilize an enzyme provides a valuable input to reduce sequence space of enzyme libraries dedicated to the optimization of functions beyond these enzyme characteristics. To allow filtering of sequence space, LibGENiE will initially attempt to predict the stability of each possible single site variant from the corresponding protein sequence employing the structure-based version of ACDC-NN, an antisymmetric neural network [63]. The structure required to run the algorithm is modeled through the ESM-esmfold_v1 API [65]. If no 3D structure of the protein of interest can be modeled, LibGENiE falls back to sequence-only predictions through ACDC-NN Seq, a model that has been described to favorably compare with other state-of-the-art sequence-based prediction tools as well as some structure-based ones [66].

3.1.2. Evolutionary information

Using the MSA, the observed conservation percentages of all 20 amino acids at each position is calculated. This information might be used to "restrict" the allowed sequence space or implement consensus/frequency ratio-based engineering techniques. The intuition behind restricting the allowed sequence space – which is to exclude residues that are never observed in closely related wild-type enzymes – is that deleterious mutations tend to be purged by natural selection [40]. Consensus or frequency ratio techniques introduce changes where the wild-type residue diverges the most from the most common amino acid (consensus) in the multiple sequence alignment. Such changes have been observed to increase stability [67–72] and are explained in detail by Damborsky et al. in their publication accompanying the release of HotSpot Wizard 2.0 [73].

3.1.3. Structural flexibility

Introducing mutations to rigidify flexible positions can yield proteins with improved stability [74]. This technique builds on the notion that selective substitutions of mobile residues can introduce additional interactions/contacts between neighbors [75,76], causing enhanced rigidity, which in turn leads to higher thermostability [77]. A typical experimental metric for protein flexibility is the B-factor, which reflects the X-ray scattering caused by thermal motion [78]. However, as B-factors are an experimental metric, and crystal structures are not available for all proteins, computational tools have been developed to predict them. In LibGENiE, we provide predictions of flexibility from one such tool, MEDUSA [64], a deep-learning-based protein flexibility model trained on experimentally determined B-factor values.

3.1.4. Oligo Design

As outlined above, oligo pools are limited in length. To enable the introduction of single point mutations at any desired position within a target sequence, the gene must consequently be split into smaller sections. Based on the provided input DNA sequence, LibGENiE's oligo design tool divides the gene into fragments of desired length including all targeted single-point mutations. In addition, the sequences of the required amplification primers are designed.

4. Conclusion

Semi-rational protein engineering is an elegant compromise between directed evolution and rational design. It directly addresses the screening bottleneck of classical directed evolution while circumventing the need to have an absolute understanding of the sequence-function relationship in enzymes (and, consequently, the required computational resources). To conduct semi-rational protein engineering, several strategies to reduce sequence space have been developed and allowed the construction of powerful enzymes for synthesis [16,22,52,79]. In this spirit, we present how the prediction and removal of destabilizing mutations in gene libraries is an effective way to reduce sequence space resulting in information-enriched gene libraries for functional screening.

However, when reducing sequence space, practical “wet-lab" experimental considerations also must be taken into account. Arbitrarily complex libraries cannot be constructed economically in most cases. Thus, improved DNA synthesis techniques will be essential to fuel the demands of an age defined by ever-increasing automation and powerful and accessible DNA sequencing instrumentation. In this vein, on-chip solid-phase gene synthesis presents itself as a compelling asset to semi-rational design as it allows to rapidly construct diverse and complex gene libraries [80]. Using this technology, researchers can build libraries tailored to their screening capabilities that can be scaled dynamically, often with no additional molecular biology overhead.

To facilitate the adoption of mutational pre-filtering, for example through the exclusion of destabilizing mutations, we introduce the webserver LibGENiE for the construction of information-enriched gene libraries. By providing data sets comprising selected common metrics used for protein engineering, LibGENiE affords researchers with a starting point for identifying hot spots and a way to restrict the sequence space to match the bounds of their screening capabilities. LibGENiE was designed to be easily extendable with additional information, whether from already available web servers for protein design such as PROSS [40], HotspotWizard [81] and 3DM [82] or other computational tools. In fact, unlike other platforms, LibGENiE provides information for all possible single-point mutants in a user's input sequence rather than suggesting preselected variants or hot spots. By providing unprocessed data, users of LibGENiE have more flexibility to introduce additional custom information and to define the number of variants to be evaluated, which can range from hundreds to thousands, depending on screening capabilities.

5. Materials and Methods:

5.1. Data

The enzyme engineering datasets used for analysis were obtained from published manuscripts [6,7,41,42]. The dataset of single mutations in ATA217 [41] was generated by extracting the 26 mutations introduced in the final variant compared to the wild-type sequence. The same procedure was applied to obtain the HG3.17 dataset [6]. The 84 beneficial mutations and their activity for the DvCA dataset were published in the supplement information of [7]. The beneficial mutations for *Aci*SHC stem from publication [42]. Beneficial single-site mutations refer to the highlighted beneficial variants obtained from a 14 single-site saturation screen (Table S1).

5.2. Cartesian ΔΔG protocol

ΔΔG predictions were based on a protocol published by the official Rosetta forums: https://www.rosettacommons.org/node/11126. Each mutant was predicted three times, and the lowest energy obtained was compared to the wild-type energies to calculate differences in free energy.

5.3. Oligo design

A pool of 200 oligo sequences with a length of <200 bp was ordered from Twist Bioscience. The sequence used were the first 157 bases of the Kemp eliminase HG3 [6]:

TGGCAGAAGCAGCACAGAGCGTTGACCAGCTGATTAAAGCACGTGGTAAAGTTTATTTTGGTGTTGCCA
CCGATCAGAATCGTCTGACCACCGGTAAAAATGCAGCAATTATTCAGGCAGATTTTGGTATGGTTTGGCC
TGAAAATAGCATGAAAT

Four distinct spatial regions along the 157 bp fragments were changed to three consecutive adenines to create diversity for analysis. Each sequence was ordered 50 times in the pool.

SeqA index: 30, 31, 32; SeqB index: 62, 63, 64; SeqC index: 93, 94, 95; SeqD index: 124, 125, 126.

The full sequences are listed in the supplementary information.

5.4. Oligo pool amplification

The oligo pools were amplified according to the protocol provided by Twist Bioscience [57]. For optimization purposes, the final dNTP concentrations (0.3 mM each dNTP or 0.6 mM each dNTP), DNA polymerase (KAPA HiFi HotStart DNA Polymerase (Roche KK2601), Q5 High-Fidelity DNA Polymerase (NEB #M0493), and Phusion High-Fidelity DNA Polymerase (NEB #M0530S) and the number of amplification cycles (15, 30, 40) were changed.

5.5. Amplified pool sequencing

After PCR amplification, the PCR pools were prepared, sequenced, and analyzed using Nanopore sequencing according to the protocol outlined in [83]. Correct sequences in which the expected nucleotide changes were detected were annotated as “1 mutation” (Figure 3c). Sequences harboring no or multiple mutations were classified as wild-type or multiple-point variants, respectively.

Author contribution

David Patsch: methodology, data collection, software implementation, analysis, writing.

Michael Eichenberger: conceptualization, methodology.

Moritz Voss: analysis, writing.

Rebecca Buller: conceptualization, methodology, writing, supervision.

Acknowledgement

This work was created as part of NCCR Catalysis, a National Centre of Competence in Research funded by the Swiss National Science Foundation (Grant number 180544).

References

[1] Schmid A, Dordick JS, Hauer B, Kiener A, Wubbolts M, Witholt B. Industrial biocatalysis today and tomorrow. Nature 2001;409:258–68. https://doi.org/10.1038/35051736.

[2] Lutz S. Beyond directed evolution-semi-rational protein engineering and design. Curr Opin Biotechnol 2010;21:734–43. https://doi.org/10.1016/j.copbio.2010.08.011.

[3] Glieder A. Protein Engineering Handbook. Edited by Stefan Lutz and Uwe T. Bornscheuer. Chembiochem 2009;10:2111–2. https://doi.org/10.1002/cbic.200900410.

[4] Büchler J, Malca SH, Patsch D, Voss M, Turner NJ, Bornscheuer UT, et al. Algorithm-aided engineering of aliphatic halogenase WelO5* for the asymmetric late-stage functionalization of soraphens. Nat Commun 2022;13. https://doi.org/10.1038/s41467-022-27999-1.

[5] Meyer F, Frey R, Ligibel M, Sager E, Schroer K, Snajdrova R, et al. Modulating Chemoselectivity in a Fe(II)/α-Ketoglutarate-Dependent Dioxygenase for the Oxidative Modification of a Nonproteinogenic Amino Acid. ACS Catal 2021;11:6261–9. https://doi.org/10.1021/acscatal.1c00678.

[6] Blomberg R, Kries H, Pinkas DM, Mittl PRE, Grütter MG, Privett HK, et al. Precision is essential for efficient catalysis in an evolved Kemp eliminase. Nature 2013;503:418–21. https://doi.org/10.1038/nature12623.

[7] Alvizo O, Nguyen LJ, Savile CK, Bresson JA, Lakhapatri SL, Solis EOP, et al. Directed evolution of an ultrastable carbonic anhydrase for highly efficient carbon capture from flue gas. Proc Natl Acad Sci U. S. A. 2014;111:16436–41. https://doi.org/10.1073/pnas.1411461111.

[8] Fox RJ, Davis SC, Mundorff EC, Newman LM, Gavrilovic V, Ma SK, et al. Improving catalytic function by ProSAR-driven enzyme evolution. Nat Biotechnol 2007;25:338–44. https://doi.org/10.1038/nbt1286.

[9] Cadet F, Fontaine N, Li G, Sanchis J, Ng Fuk Chong M, Pandjaitan R, et al. A machine learning approach for reliable prediction of amino acid interactions and its application in the directed evolution of enantioselective enzymes. Sci Rep 2018;8. https://doi.org/10.1038/s41598-018-35033-y.

[10] Liang F, Feng XJ, Lowry M, Rabitz H. Maximal use of minimal libraries through the adaptive substituent reordering algorithm. J Phys Chem B 2005;109:5842–54. https://doi.org/10.1021/jp045926y.

[11] Turner NJ. Directed evolution drives the next generation of biocatalysts. Nat Chem Biol 2009;5:567–73. https://doi.org/10.1038/nchembio.203.

[12] Reetz MT, Kahakeaw D, Lohmer R. Addressing the numbers problem in directed evolution. ChemBioChem 2008;9:1797–804. https://doi.org/10.1002/cbic.200800298.

[13] Balke K, Beier A, Bornscheuer UT. Hot spots for the protein engineering of Baeyer-Villiger monooxygenases. Biotechnol Adv 2018;36:247–63. https://doi.org/10.1016/j.biotechadv.2017.11.007.

[14] Reetz MT, Wang LW, Bocola M. Directed evolution of enantioselective enzymes: Iterative cycles of CASTing for probing protein-sequence space. Angew Chem Int Ed 2006;45:1236–41. https://doi.org/10.1002/anie.200502746.

[15] Bloom JD, Labthavikul ST, Otey CR, Arnold FH. Protein stability promotes evolvability. Proc Natl Acad Sci U.S.A. 2006;103:5869–74. https://doi.org/10.1073/pnas.0510098103.

[16] Reetz M. Making Enzymes Suitable for Organic Chemistry by Rational Protein Design. ChemBioChem 2022. https://doi.org/10.1002/cbic.202200049.

[17] Kazlauskas R, Bornscheuer U. Finding better protein engineering strategies. Nat Chem Biol 2009;5:526–9. https://doi.org/10.1038/nchembio0809-526.

[18] Jumper J, Evans R, Pritzel A, Green T, Figurnov M, Ronneberger O, et al. Highly accurate protein structure prediction with AlphaFold. Nature 2021;596:583–9. https://doi.org/10.1038/s41586-021-03819-2.

[19] Mehmood R, Vennelakanti V, Kulik HJ. Revealing Substrate Positioning Dynamics in Non-heme Fe(II)/αKG-dependent Halogenases Through Spectroscopically Guided Simulation. ChemRxiv 2021. https://doi.org/10.26434/chemrxiv-2021-m7dh3.

[20] Porebski BT, Buckle AM. Consensus protein design. Protein Eng Des Sel 2016;29:245–51. https://doi.org/10.1093/protein/gzw015.

[21] Kaushik M, Sinha P, Jaiswal P, Mahendru S, Roy K, Kukreti S. Protein engineering and de novo designing of a biocatalyst. J Mol Recognit 2016;29:499–503. https://doi.org/10.1002/jmr.2546.

[22] Reetz MT. Laboratory evolution of stereoselective enzymes: A prolific source of catalysts for asymmetric reactions. Angew Chem Int Ed 2011;50:138–74. https://doi.org/10.1002/anie.201000826.

[23] Reetz MT. Biocatalysis in organic chemistry and biotechnology: Past, present, and future. J Am Chem Soc 2013;135:12480–96. https://doi.org/10.1021/ja405051f.

[24] Wilding M, Hong N, Spence M, Buckle AM, Jackson CJ. Protein engineering: The potential of remote mutations. Biochem Soc Trans 2019;47:701–11. https://doi.org/10.1042/BST20180614.

[25] Li D, Wu Q, Reetz MT. Focused rational iterative site-specific mutagenesis (FRISM). Methods Enzymol, vol. 643, Academic Press Inc.; 2020, p. 225–42. https://doi.org/10.1016/bs.mie.2020.04.055.

[26] Beaucage SL, Caruthers MH. Deoxynucleoside phosphoramidites—A new class of key intermediates for deoxypolynucleotide synthesis. vol. 22. 1981. https://doi.org/https://doi.org/10.1016/S0040-4039(01)90461-7.

[27] Kosuri S, Church GM. Large-scale de novo DNA synthesis: Technologies and applications. Nat Methods 2014;11:499–507. https://doi.org/10.1038/nmeth.2918.

[28] Oligo pool pricing - twist n.d. https://ecommerce.twistdna.com/app/oligo (accessed April 26, 2023).

[29] Kuiper BP, Prins RC, Billerbeck S. Oligo Pools as an Affordable Source of Synthetic DNA for Cost-Effective Library Construction in Protein- and Metabolic Pathway Engineering. ChemBioChem 2022;23. https://doi.org/10.1002/cbic.202100507.

[30] Victorino da Silva Amatto I, Gonsales da Rosa-Garzon N, Antônio de Oliveira Simões F, Santiago F, Pereira da Silva Leite N, Raspante Martins J, et al. Enzyme engineering and its industrial applications. Biotechnol Appl Biochem 2022;69:389–409. https://doi.org/10.1002/bab.2117.

[31] Reetz MT, Wang LW, Bocola M. Directed evolution of enantioselective enzymes: Iterative cycles of CASTing for probing protein-sequence space. Angew Chem Int Ed 2006;45:1236–41. https://doi.org/10.1002/anie.200502746.

[32] Childers MC, Daggett V. Insights from molecular dynamics simulations for computational protein design. Mol Syst Des Eng 2017;2:9–33. https://doi.org/10.1039/c6me00083e.

[33] Cao H, Wang J, He L, Qi Y, Zhang JZ. DeepDDG: Predicting the Stability Change of Protein Point Mutations Using Neural Networks. J Chem Inf Model 2019;59:1508–14. https://doi.org/10.1021/acs.jcim.8b00697.

[34] Quan L, Lv Q, Zhang Y. STRUM: Structure-based prediction of protein stability changes upon single-point mutation. Bioinformatics 2016;32:2936–46. https://doi.org/10.1093/bioinformatics/btw361.

[35] Kellogg EH, Leaver-Fay A, Baker D. Role of conformational sampling in computing mutation-induced changes in protein structure and stability. Proteins 2011;79:830–8. https://doi.org/10.1002/prot.22921.

[36] Alford RF, Leaver-Fay A, Jeliazkov JR, O'Meara MJ, DiMaio FP, Park H, et al. The Rosetta All-Atom Energy Function for Macromolecular Modeling and Design. J Chem Theory Comput 2017;13:3031–48. https://doi.org/10.1021/acs.jctc.7b00125.

[37] Giollo M, Martin AJM, Walsh I, Ferrari C, Tosatto SCE. NeEMO: a method using residue interaction networks to improve prediction of protein stability upon mutation. BMC Genomics 2014. https://doi.org/https://doi.org/10.1186/1471-2164-15-S4-S7.

[38] Chen C-W, Lin J, Chu Y-W. iStable: off-the-shelf predictor integration for predicting protein stability changes. BMC Bioinformatics 2013;14:S5. https://doi.org/10.1186/1471-2105-14-S2-S5.

[39] Usmanova DR, Bogatyreva NS, Bernad JA, Eremina AA, Gorshkova AA, Kanevskiy GM, et al. Self-consistency test reveals systematic bias in programs for prediction change of stability upon mutation. Bioinformatics 2018;34:3653–8. https://doi.org/10.1093/bioinformatics/bty340.

[40] Goldenzweig A, Goldsmith M, Hill SE, Gertman O, Laurino P, Ashani Y, et al. Automated Structure- and Sequence-Based Design of Proteins for High Bacterial Expression and Stability. Mol Cell 2016;63:337–46. https://doi.org/10.1016/j.molcel.2016.06.012.

[41] Novick SJ, Dellas N, Garcia R, Ching C, Bautista A, Homan D, et al. Engineering an amine transaminase for the efficient production of a chiral sacubitril precursor. ACS Catal 2021;11:3762–70. https://doi.org/10.1021/acscatal.0c05450.

[42] Eichenberger M, Hüppi S, Patsch D, Aeberli N, Berweger R, Dossenbach S, et al. Asymmetric Cation-Olefin Monocyclization by Engineered Squalene–Hopene Cyclases. Angew Chem Int Ed 2021;60:26080–6. https://doi.org/10.1002/anie.202108037.

[43] Kellogg EH, Leaver-Fay A, Baker D. Role of conformational sampling in computing mutation-induced changes in protein structure and stability. Proteins 2011;79:830–8. https://doi.org/10.1002/prot.22921.

[44] Frenz B, Lewis SM, King I, DiMaio F, Park H, Song Y. Prediction of Protein Mutational Free Energy: Benchmark and Sampling Improvements Increase Classification Accuracy. Front Bioeng Biotechnol 2020;8. https://doi.org/10.3389/fbioe.2020.558247.

[45] Jomain JB, Tallet E, Broutin I, Hoos S, van Agthoven J, Ducruix A, et al. Structural and thermodynamic bases for the design of pure prolactin receptor antagonists: X-ray structure of Del1-9-G129R-hPRL. J Biol Chem 2007;282:33118–31. https://doi.org/10.1074/jbc.M704364200.

[46] Torrado M, Revuelta J, Gonzalez C, Corzana F, Bastida A, Asensio JL. Role of conserved salt bridges in homeodomain stability and DNA binding. J Biol Chem 2009;284:23765–79. https://doi.org/10.1074/jbc.M109.012054.

[47] Yokota A, Takahashi H, Takenawa T, Arai M. Probing the roles of conserved arginine-44 of Escherichia coli dihydrofolate reductase in its function and stability by systematic sequence perturbation analysis. Biochem Biophys Res Commun 2010;391:1703–7. https://doi.org/10.1016/j.bbrc.2009.12.134.

[48] Fredricksen RS, Swenson CA. Relationship between stability and function for isolated domains of troponin C. Biochemistry 1996;35:14012–26. https://doi.org/1021/bi961270q.

[49] Zakrzewska M, Krowarsch D, Wiedlocha A, Olsnes S, Otlewski J. Highly stable mutants of human fibroblast growth factor-1 exhibit prolonged biological action. J Mol Biol 2005;352:860–75. https://doi.org/10.1016/j.jmb.2005.07.066.

[50] Kragelund BB, Jönsson M, Bifulco G, Chazin WJ, Nilsson H, Finn BE, et al. Hydrophobic Core Substitutions in Calbindin D9k: Effects on Ca2+ Binding and Dissociation. Biochemistry 1998;37:8926–37. https://doi.org/10.1021/bi9726436.

[51] Chaparro-Riggers JF, Polizzi KM, Bommarius AS. Better library design: Data-driven protein engineering. Biotechnol J 2007;2:180–91. https://doi.org/10.1002/biot.200600170.

[52] Reetz MT, Wu S. Greatly reduced amino acid alphabets in directed evolution: Making the right choice for saturation mutagenesis at homologous enzyme positions. ChemComm 2008:5499–501. https://doi.org/10.1039/b813388c.

[53] Horton RM, Cai Z, Ho SN, Pease LR. Gene Splicing by Overlap Extension: Tailor-Made Genes Using the Polymerase Chain Reaction. BioTechniques 2013;54:129–33. https://doi.org/10.2144/000114017.

[54] Faber MS, Van Leuven JT, Ederer MM, Sapozhnikov Y, Wilson ZL, Wichman HA, et al. Saturation mutagenesis genome engineering of infective ϕx174 bacteriophage via unamplified oligo pools and golden gate assembly. ACS Synth Biol 2020;9:125–31. https://doi.org/10.1021/acssynbio.9b00411.

[55] Kitzman JO, Starita LM, Lo RS, Fields S, Shendure J. Massively parallel single-amino-acid mutagenesis. Nat Methods 2015;12:203–6. https://doi.org/10.1038/nmeth.3223.

[56] Steiner P, Baumer Z, Whitehead T. A Method for User-defined Mutagenesis by Integrating Oligo Pool Synthesis Technology with Nicking Mutagenesis. Bio Protoc 2020;10. https://doi.org/10.21769/bioprotoc.3697.

[57] twist-oligo-pool-amplification-guidelines. https://www.twistbioscience.com/resources/protocol/twist-oligo-pool-amplification-guidelines (accessed April 26, 2023).

[58] Becker M, Noll-Puchta H, Amend D, Nolte F, Fuchs C, Jeremias I, et al. CLUE: A bioinformatic and wet-lab pipeline for multiplexed cloning of custom sgRNA libraries. Nucleic Acids Res 2020;48. https://doi.org/10.1093/nar/gkaa459.

[59] Meyerhans A, Vartanian J-P, Wain-Hobson S. DNA recombination during PCR. Nucleic Acids Res 1990;18:1687–91. https://doi.org/10.1093/nar/18.7.1687.

[60] Judo MS, Wedel AB, Wilson C. Stimulation and suppression of PCR-mediated recombination. Nucleic Acids Res 1998;26:1819–25. https://doi.org/10.1093/nar/26.7.1819.

[61] Hegde M, Strand C, Hanna RE, Doench JG. Uncoupling of sgRNAs from their associated barcodes during PCR amplification of combinatorial CRISPR screens. PLoS One 2018;13. https://doi.org/10.1371/journal.pone.0197547.

[62] Madeira F, Park YM, Lee J, Buso N, Gur T, Madhusoodanan N, et al. The EMBL-EBI search and sequence analysis tools APIs in 2019. Nucleic Acids Res 2019;47:W636–41. https://doi.org/10.1093/nar/gkz268.

[63] Benevenuta S, Pancotti C, Fariselli P, Birolo G, Sanavia T. An antisymmetric neural network to predict free energy changes in protein variants. J Phys D Appl Phys 2021;54. https://doi.org/10.1088/1361-6463/abedfb.

[64] Vander Meersche Y, Cretin G, de Brevern AG, Gelly JC, Galochkina T. MEDUSA: Prediction of Protein Flexibility from Sequence. J Mol Biol 2021;433. https://doi.org/10.1016/j.jmb.2021.166882.

[65] Lin Z, Akin H, Rao R, Hie B, Zhu Z, Lu W, et al. Evolutionary-scale prediction of atomic level protein structure with a language model 2021. https://doi.org/10.1101/2022.07.20.500902.

[66] Pancotti C, Benevenuta S, Repetto V, Birolo G, Capriotti E, Sanavia T, et al. A deep-learning sequence-based method to predict protein stability changes upon genetic variations. Genes 2021;12. https://doi.org/10.3390/genes12060911.

[67] Amin N, Liu AD, Ramer S, Aehle W, Meijer D, Metin M, et al. Construction of stabilized proteins by combinatorial consensus mutagenesis. Protein Eng Del Sel 2004;17:787–93. https://doi.org/10.1093/protein/gzh091.

[68] Pey AL, Rodriguez-Larrea D, Bomke S, Dammers S, Godoy-Ruiz R, Garcia-Mira MM, et al. Engineering proteins with tunable thermodynamic and kinetic stabilities. Proteins 2008;71:165–74. https://doi.org/10.1002/prot.21670.

[69] Sullivan BJ, Nguyen T, Durani V, Mathur D, Rojas S, Thomas M, et al. Stabilizing proteins from sequence statistics: The interplay of conservation and correlation in triosephosphate isomerase stability. J Mol Biol 2012;420:384–99. https://doi.org/10.1016/j.jmb.2012.04.025.

[70] Magliery TJ. Protein stability: Computation, sequence statistics, and new experimental methods. Curr Opin Struct Biol 2015;33:161–8. https://doi.org/10.1016/j.sbi.2015.09.002.

[71] Steipe B, Schiller B, Plückthun A, Steinbacher S. Sequence Statistics Reliably Predict Stabilizing Mutations in a Protein Domain. J Mol Biol 1994;240:188–92. https://doi.org/https://doi.org/10.1006/jmbi.1994.1434.

[72] Lehmann M, Loch C, Middendorf A, Studer D, Lassen SF, Pasamontes L, et al. The consensus concept for thermostability engineering of proteins: further proof of concept. Protein Eng Des Sel 2002;15:403–11. https://doi.org/10.1093/protein/15.5.403.

[73] Bendl J, Stourac J, Sebestova E, Vavra O, Musil M, Brezovsky J, et al. HotSpot Wizard 2.0: automated design of site-specific mutations and smart libraries in protein engineering. Nucleic Acids Res 2016;44:W479–87. https://doi.org/10.1093/nar/gkw416.

[74] Yu H, Huang H. Engineering proteins for thermostability through rigidifying flexible sites. Biotechnol Adv 2014;32:308–15. https://doi.org/10.1016/j.biotechadv.2013.10.012.

[75] Jochens H, Aerts D, Bornscheuer UT. Thermostabilization of an esterase by alignment-guided focussed directed evolution. Protein Eng Des Sel 2010;23:903–9. https://doi.org/10.1093/protein/gzq071.

[76] Cerdobbel A, de Winter K, Aerts D, Kuipers R, Joosten HJ, Soetaert W, et al. Increasing the thermostability of sucrose phosphorylase by a combination of sequence- and structure-based mutagenesis. Protein Eng Des Sel 2011;24:829–34. https://doi.org/10.1093/protein/gzr042.

[77] Reetz MT, Soni P, Fernández L, Gumulya Y, Carballeira JD. Increasing the stability of an enzyme toward hostile organic solvents by directed evolution based on iterative saturation mutagenesis using the B-FIT method. ChemComm 2010;46:8657–8. https://doi.org/10.1039/c0cc02657c.

[78] Sun Z, Liu Q, Qu G, Feng Y, Reetz MT. Utility of B-Factors in Protein Science: Interpreting Rigidity, Flexibility, and Internal Motion and Engineering Thermostability. Chem Rev 2019. https://doi.org/10.1021/acs.chemrev.8b00290.

[79] Reetz MT, Carballeira JD. Iterative saturation mutagenesis (ISM) for rapid directed evolution of functional enzymes. Nat Protoc 2007;2:891–903. https://doi.org/10.1038/nprot.2007.72.

[80] Qu G, Li A, Acevedo-Rocha CG, Sun Z, Reetz MT. The Crucial Role of Methodology Development in Directed Evolution of Selective Enzymes. Angew Chem Int Ed 2020;59:13204–31. https://doi.org/10.1002/anie.201901491.

[81] Sumbalova L, Stourac J, Martinek T, Bednar D, Damborsky J. HotSpot Wizard 3.0: Web server for automated design of mutations and smart libraries based on sequence input information. Nucleic Acids Res 2018;46:W356–62. https://doi.org/10.1093/nar/gky417.

[82] Kuipers RK, Joosten HJ, Van Berkel WJH, Leferink NGH, Rooijen E, Ittmann E, et al. 3DM: Systematic analysis of heterogeneous superfamily data to discover protein functionalities. Proteins 2010;78:2101–13. https://doi.org/10.1002/prot.22725.

[83] Currin A, Swainston N, Dunstan MS, Jervis AJ, Mulherin P, Robinson CJ, et al. Highly multiplexed, fast and accurate nanopore sequencing for verification of synthetic DNA constructs and sequence libraries ACS Synth Biol. https://doi.org/10.1093/synbio/ysz025.

Supporting Information

LibGENiE – A bioinformatic pipeline for the design of information-enriched enzyme libraries

David Patsch[a,b], Michael Eichenberger[a], Moritz Voss[a], Uwe T. Bornscheuer[b] and Rebecca M. Buller[a,*]

[a] Zurich University of Applied Sciences, School of Life Sciences and Facility Management, Institute of Chemistry and Biotechnology, Einsiedlerstrasse 31, 8820 Wädenswil, Switzerland

[b] Institute of Biochemistry, Department of Biotechnology & Enzyme Catalysis, Greifswald University, Felix-Hausdorff-Strasse 4, 17487 Greifswald, Germany

*Corresponding author: Rebecca M. Buller (rebecca.buller@zhaw.ch)

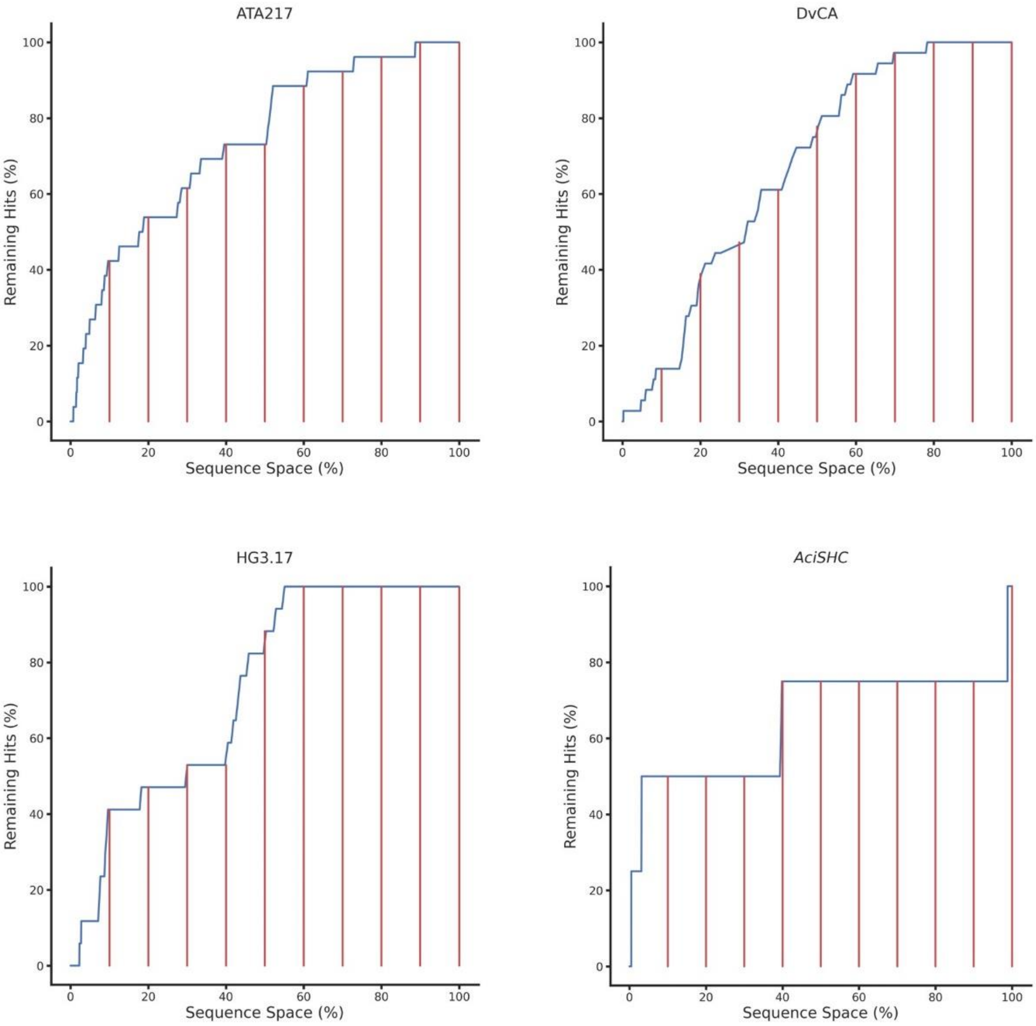


Figure S1: Visual representation of how reductions in sequence space (filtered by ddG) would affect the number of hits identified in the evolutionary trajectories of a transaminase (ATA217) [1], a carbonic anhydrase (DvCA) [2], a computationally designed Kemp Eliminase (HG3.17) [3] and a squalene-hopene cyclase (*Aci*SHC) [4]. The underlying data can be found in Table 1. Vertical red lines are drawn in 10 % increments.

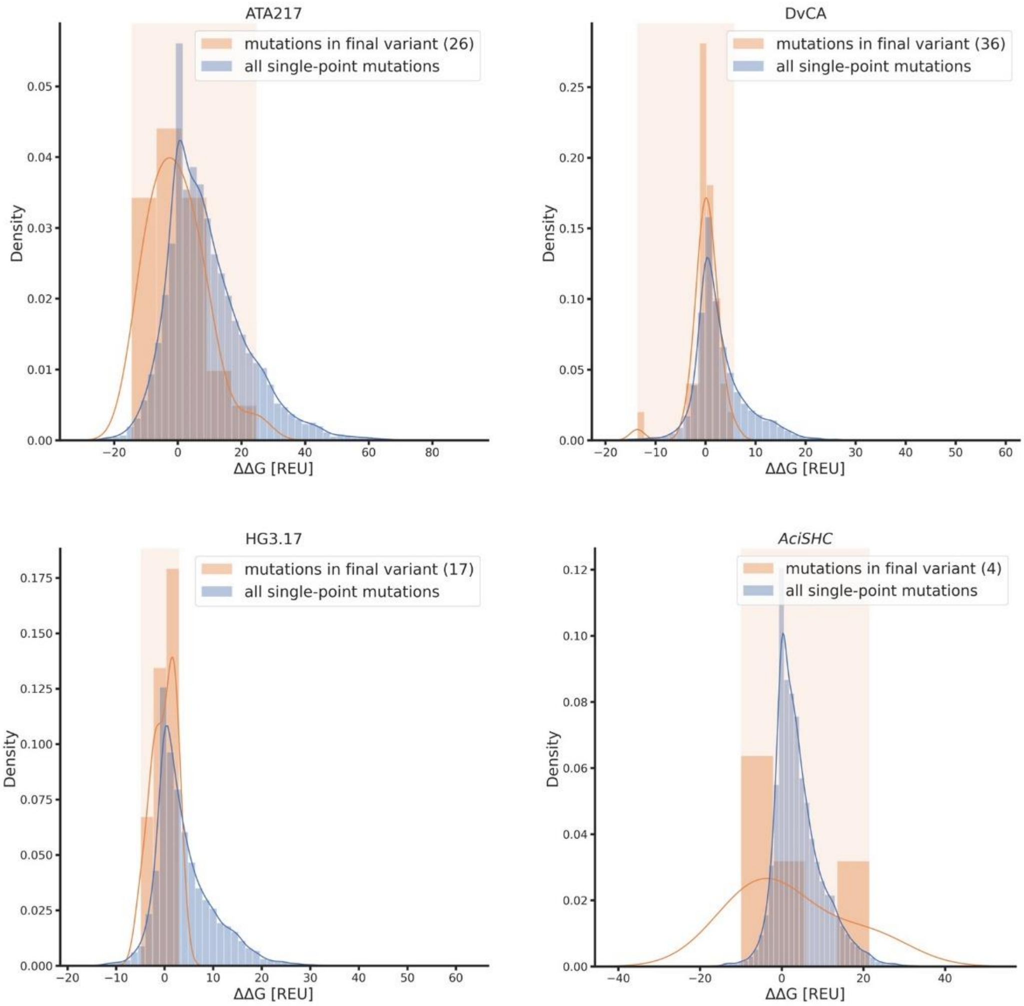


Figure S2: Density plot of predicted ddG values (lower values correspond to higher predicted stability) of transaminase (ATA217) [1], a carbonic anhydrase [2], a computationally designed Kemp Eliminase (HG3.17) [3] and a squalene-hopene cyclase (*Aci*SHC) [4]. The blue density curve depicts the ddG values of all possible single-point mutants, and the orange plot represents the ddG distribution of the reported hits. The ddG range in which hits were identified is highlighted in orange.

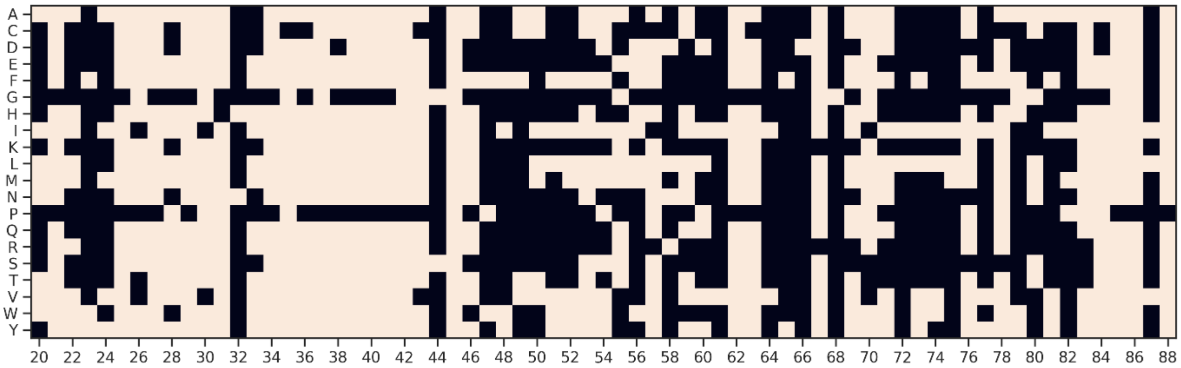


Figure S3: Amino acid distribution of the DvCA gene after removing the 40 % mutants with the highest predicted ddG. Removed residues are colored in black.

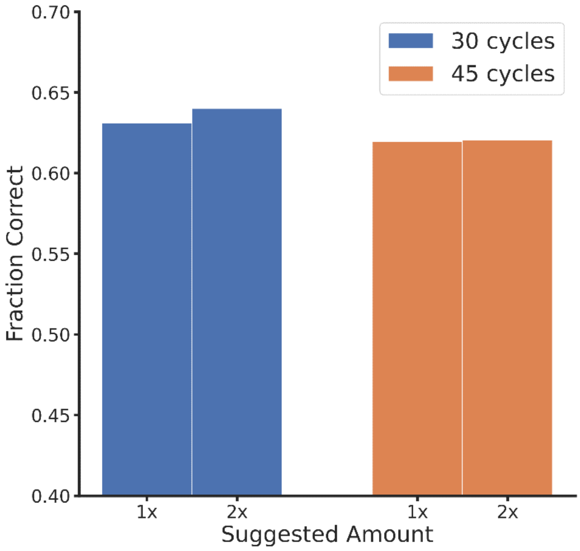


Figure S4: Impact of dNTP concentration on the number of correct sequences for experiments conducted with KAPA polymerase at 30 and 45 cycles. The suggested amount refers to 0.3 mM of each dNTP in the amplification reaction according to the PCR AMPLIFICATION PROTOCOL provided by Twist.

Sequence space (%)

	# mut	100.0	90	80	70	60	50	40	30	20	10
ATA217	**26**	100.0	100.0	96.2	92.3	88.5	73.1	73.1	61.5	53.8	42.3
HG3.17	**17**	100.0	100.0	100.0	100.0	100.0	82.4	52.9	52.9	47.1	41.2
DvCA_Final	**36**	100.0	100.0	100.0	97.2	91.7	77.8	61.1	44.4	38.9	13.9
DvCA_Single	**84**	100.0	100.0	100.0	97.6	92.9	77.4	57.1	39.3	33.3	13.1
***Aci*SHC_Final**	**4**	100.0	75.0	75.0	75.0	75.0	75.0	75.0	50.0	50.0	50.0
***Aci*SHC_Single**	**9**	100.0	100.0	88.9	88.9	77.8	77.8	55.6	44.4	44.4	11.1
Average		100.0	95.8	93.3	91.8	87.6	77.2	62.5	48.8	44.6	28.6

Table S1: Extended overview of different evolution campaigns and how improved variants are distributed concerning ddG. This table includes the additional entries DvCA_Single and *Aci*SHC_Single. DvCA_Single refers to the 84 single-point mutations obtained during the NNK screening (as described in the main text). The *Aci*SHC_Single row refers to the nine single-point mutations initially identified as beneficial of which only four were included in the final SHC variant.

Sequences:

> sequence A

TAATACGACTCACTATAGGGATGGCAGAAGCAGCACAGAGCGTTGACCAGCAAATTAAAGCACGTGGTAAAGTTTATTTTGGTGTTGCCACCGATCAGAATCGTCTGACCACCGGTAAAAATGCAGCAATTATTCAGGCAGATTTTGGTATGGTTTGGCCTGAAAATAGCATGAACTGAGCAATAACTAGCATAA

> sequence B

TAATACGACTCACTATAGGGATGGCAGAAGCAGCACAGAGCGTTGACCAGCTGATTAAAGCACGTGGTAAAGTTTATTTTGGAAATGCCACCGATCAGAATCGTCTGACCACCGGTAAAAATGCAGCAATTATTCAGGCAGATTTTGGTATGGTTTGGCCTGAAAATAGCATGAACTGAGCAATAACTAGCATAA

> sequence C

TAATACGACTCACTATAGGGATGGCAGAAGCAGCACAGAGCGTTGACCAGCTGATTAAAGCACGTGGTAAAGTTTATTTTGGTGTTGCCACCGATCAGAATCGTCTGACCACCAAAAAAAATGCAGCAATTATTCAGGCAGATTTTGGTATGGTTTGGCCTGAAAATAGCATGAACTGAGCAATAACTAGCATAA

> sequence D

TAATACGACTCACTATAGGGATGGCAGAAGCAGCACAGAGCGTTGACCAGCTGATTAAAGCACGTGGTAAAGTTTATTTTGGTGTTGCCACCGATCAGAATCGTCTGACCACCGGTAAAAATGCAGCAATTATTCAGGCAGATTAAAGTATGGTTTGGCCTGAAAATAGCATGAACTGAGCAATAACTAGCATAA

References

[1] Novick SJ, Dellas N, Garcia R, Ching C, Bautista A, Homan D, et al. Engineering an amine transaminase for the efficient production of a chiral sacubitril precursor. ACS Catal 2021;11:3762–70. https://doi.org/10.1021/acscatal.0c05450.

[2] Alvizo O, Nguyen LJ, Savile CK, Bresson JA, Lakhapatri SL, Solis EOP, et al. Directed evolution of an ultrastable carbonic anhydrase for highly efficient carbon capture from flue gas. Proc Natl Acad Sci U. S. A. 2014;111:16436–41. https://doi.org/10.1073/pnas.1411461111.

[3] Blomberg R, Kries H, Pinkas DM, Mittl PRE, Grütter MG, Privett HK, et al. Precision is essential for efficient catalysis in an evolved Kemp eliminase. Nature 2013;503:418–21. https://doi.org/10.1038/nature12623.

[4] Eichenberger M, Hüppi S, Patsch D, Aeberli N, Berweger R, Dossenbach S, et al. Asymmetric Cation-Olefin Monocyclization by Engineered Squalene–Hopene Cyclases. Angew Chem Int Ed 2021;60:26080–6. https://doi.org/10.1002/anie.202108037.

Article IV

Draft: Application of the filter/oligo methodology for the re-evolution of the Kemp eliminase HG3.

David Patsch[a,b], Moritz Voss[a], Uwe T. Bornscheuer[b] and Rebecca M. Buller[a, *]

[a] Zurich University of Applied Sciences, School of Life Sciences and Facility Management, Institute of Chemistry and Biotechnology, Einsiedlerstrasse 31, 8820 Wädenswil, Switzerland

[b] Institute of Biochemistry, Department of Biotechnology & Enzyme Catalysis, Greifswald University, Felix-Hausdorff-Strasse 4, D17487 Greifswald, Germany

An early draft manuscript summarizing the evolutionary trajectory and experimental design. This manuscript is limited to the work carried out by David Patsch and Moritz Voss and builds heavily on the strategy/methodology outlined in **Article III**. We intend the final manuscript to be published after the publication of **Article III**.

Introduction

The application of enzymes has seen a sharp increase over the past few decades, as their ability to catalyze reactions with exquisite selectivity and a high substrate specificity make them attractive candidates for industrial applications [1], [2]. For example, nitrile hydratases that can selectively hydrolyse a nitrile to the amide are employed in the 650 000 t / year production of acrylamide from acrylonitrile [3], [4]. Enzymes also play a role in the large-scale (10^7 tons per year) production of high-fructose corn syrup from glucose [4], [5]. Additionally, the use of biocatalysts as catalysts for synthesizing complex molecules, particularly pharmaceuticals, has recently gained widespread popularity [6]. However, the catalysts produced by nature are often unsatisfactory to perform such tasks at an industrial scale and require additional reconfiguration and optimization. For this reason, researchers rely on protein engineering techniques such as directed evolution or rational design [7]. Today, these methods are routinely applied and have resulted in astounding improvements in various protein characteristics, such as activity [8]–[10], stereoselectivity [11], [12], thermostability [13], and solvent tolerance [14].

A landmark example of the potential of protein engineering is the HG series of kemp eliminases. In this particular case, the kemp elimination (KE) reaction (Scheme 1) is not even catalyzed by naturally occurring enzymes [15], [16]. As such, the initial starting point had to be designed computationally. Based on quantum mechanical transition state calculations, an idealized active pocket was designed to stabilize the targeted transition state [16], [17]. Initially, the design HG1 did not result in any measurable KE activity. However, the information acquired throughout this process was enough to adjust the design in subsequent rounds, resulting in the designed Kemp Eliminase, dubbed HG3 [17]. As the performance of this artificial enzyme was still far below that of natural enzymes, it was subjected to multiple rounds of directed evolution based on an optimization strategy focused on global and local mutagenesis. First, error-prone PCR and gene shuffling were employed to identify regions of interest, which were then further investigated through focused libraries [8]. Finally, after 17 rounds of evolution, this resulted in a novel variant, HG3.17 containing 17 mutations, which possessed a significantly increased activity, thermostability, and protein yield [8].

The KE reaction and the HG3 series have been studied extensively and served as a model system for many projects [8], [16]–[20]. Here, we are particularly interested in the protein engineering approach used to evolve HG3 to HG3.17. Directed evolution, pioneered in the 1990s by Arnold and coworkers [21], is a fascinating concept that mimics natural evolution in the laboratory by undergoing iterative cycles of gene diversification and selection of protein variants. However, unlike nature, which primarily selects for reproduction or survival, directed evolution can be used to precisely alter properties of interest [22], creating enzyme lineages that exhibit improved or novel functionalities [22]. Diversity in traditional directed evolution is usually generated through either random PCR or gene shuffling [23]. This mostly leads to the random incorporation of mutations, which can be beneficial, as it might reveal interesting hot spots anywhere on the protein [24]. However, this also leads to much redundancy during screening, as most mutations are either neutral or unfavorable [25].

We analyzed the advantages and disadvantages of protein engineering strategies in a recent study [24] and explored how advances in computational design and gene synthesis could allow for more efficient ways of protein engineering. Here, we applied these findings to re-evolve the Kemp eliminase HG3. We reason that removing specific undesirable mutations (filtering the sequence space) and relying on the commercially available "oligo pools" for library construction can constitute a robust new methodology in the toolbox of a protein engineer. Interestingly, our findings reinforce our initial hypothesis and reveal a completely distinct evolutionary trajectory of HG3, helping us understand more of the vast protein landscape.

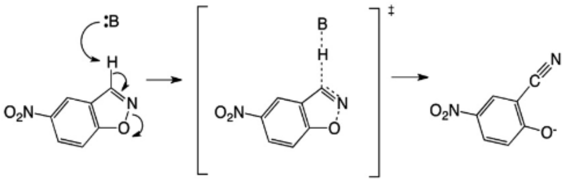


Fig. 1. The KE reaction scheme.

Scheme 1: The kemp elimination reaction in HG3.17 proceeds by deprotonating 5-nitrobenzisoxazole (1), yielding salicylonitrile (3). The base (Asp127) deprotonates 1, and the H-bond donor (Gln50) stabilizes the partial negative charge on the phenolic oxygen at the transition state (‡). Image from [16].

Results

Library Design

The HG3.17 variant differs from the starting point HG3 by 17 mutations. Nine of these mutations are situated in the immediate vicinity (<8 Å) of the co-crystallized transition state analogue 6-nitrobenzotriazole (6NT) or the tunnel leading to the active pocket [8]. The remaining eight mutations are spread across the protein; some are up to 22 Å away from the active pocket. The most critical residue identified over the evolutionary trajectory of HG3 is K50Q. The site was initially mutated to histidine in variant HG3.3 (K50H) and then again to glutamine in HG3.7 (K50Q) - it is hypothesized that this mutation can stabilize the negative charge developing on the phenolic oxygen at the transition state [8], [19]. Notably, not all mutations have the same impact on activity. A recent follow-up study, based on X-ray crystallography and computational analysis, shows that eight out of the 17 mutations in HG3.17 are enough to reach ~75 % of its activity [19]. In our previous work, we reason that such rationalizations are hard to perform *a priori* [24]. Selecting specific sites and residues to mutate to improve a desired trait is a challenging aspect of rational design. Predicting mutations that hinder expression or folding, however, is something that we can do much more reliably. None of the 17 mutations introduced into HG3.17 are predicted to destabilize the protein, indicating that a large part of the sequence space could have been excluded while still reaching the same result. Specifically, more than 40 % of all mutations could have been removed (based on stability calculations) without sacrificing a single hit [24].

The observation that beneficial mutations tend not to be enormously destabilizing could serve as the foundation for a powerful protein engineering technique that can generalize to various objectives rather than developing various strategies for each target. Identifying destabilizing mutations can be achieved through multiple means. The most common methods are computational mutation scans and aggregating evolutionary information [24]. We relied on these techniques to filter the sequence space of HG3 to roughly 1800 variants. To avoid bias and the possibility of being influenced by previous knowledge, the residues in an 8 Å radius of 6NT and the tunnel leading there are not pre-filtered but rather saturated fully (~35 sites). We selected residues from the evolutionary space and a computational single-site saturation scan for the remainder. The decision to reduce the sequence space to 1800 variants was based on multiple factors—primarily the size of the purchasable oligo pools and practical considerations regarding screening and throughput. We aimed for approximately 65 % library coverage, in line with reports from Codexis [26], [27]. Effectively, this strategy resulted in a screening burden of roughly 20 plates per round. However, as oligo pools suffer from relatively high error rates [28]–[30], we estimate the effective amount of unique variants measured per round to be closer to 600 rather than the targeted 1100. This initial screening resulted in 5-10 hits (variants with an improved activity compared to the parent), which we subsequently combined in combinatorial libraries[26], [27]. However, rather than performing statistical analysis on combinatorial variants, we selected the best-performing variant in this library and moved to the next round. This was partly because the relatively small amount of hits allowed for exhaustive screening of the combinatorial library but also because this project focused on exploring the practical implications of the filter/oligo approach.

Evolutionary trajectory

The process of filtering the sequence space to residues in the active site, tunnels, and variants identified from our computational analysis, then screening these mutations and combining hits in small combinatorial libraries, was performed five times. 2500 and 3000 variants were screened for each round, split between the initial hit identification and subsequent hit combination. Notably, after five rounds of evolution, our final variant HG3.R5 showed similar activity to HG3.17 under assay conditions. A total of 16 new mutations were introduced into HG3.R5 compared to the wild type, contrasted with the 17 in HG3.17. Astoundingly, HG3.R5, and HG3.17 only share one mutation: K50Q, which was identified as critical for catalysis [8], [19].

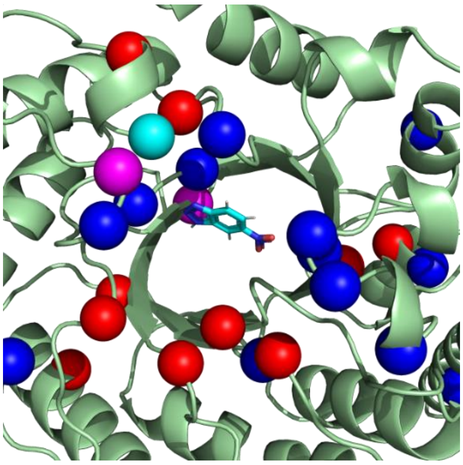


Figure 1: Comparison of the mutations of HG3.R5 and HG3.17. Mutations in the final HG3.R5 variant are colored in red. The mutations corresponding to HG3.17 are colored in blue. K50Q, the mutation that both variants have in common, is highlighted in turquoise. Pink indicates the sites that both variants substitute, though with different residues.

Discussion

While the diverging evolutionary trajectories between HG3.R5 and HG3.17 are fascinating, this manuscript focuses on the practical application of the filter/oligo methodology. Directed evolution is still a complex process, and researchers have to rely on various techniques to approach different problems. This creates significant overhead and inevitably leads to inefficiencies. For example, if the goal is to improve the thermostability of a protein, a good starting point could be the Protein Repair One-Stop Shop (PROSS) [31], a web server that allows researchers to design enzyme variants that are predicted to be more stable. The results PROSS provides are ~10 gene sequences containing an increasing number of mutations. However, the algorithm that predicts these variants is not publicly accessible, making PROSS hard to extend. The tool raises additional questions. For example, what if resources to order much more than ten genes are available? What if we have a system with a straightforward screening system (such as is the case with HG3)? There could be advances in stability predictions, which might not be updated in the online tool; how do we adjust to that? In addition, even assuming perfect predictive accuracy, the rigidification/stabilization of our enzyme of interest could decrease activity [32]–[37], a relationship that is hard to generalize. With this, we want to highlight that even for a single enzyme property, in this case, thermostability, many different aspects have to be considered, and outlining an optimal engineering strategy can be far from trivial. A typical evolution project already has many moving parts, with different goals and objectives that must be balanced. As such, the focus of the oligo/filter methodology is not necessarily to find the best solution –but rather one that is good enough for many applications. We reason that mutations that lead to misfolded, poorly expressed, or unfolded proteins are unlikely to improve a desired enzyme function.

Consequently, irrespective of the objective, a researcher can reach to the same methodology, allowing for much faster iterations and shorter rounds of evolution. This generalization level also means it is possible to automate specific processes.[24] outlines a web server that helps users find destabilizing mutations and design oligo pools to construct their libraries, which users can freely extend because it provides raw results rather than processed predictions. Depending on the project and licenses, different tools to predict undesired sequences might be required; however, these can be used interchangeably without affecting the underlying process, making adopting new and improved tools easy.

As the theoretical sequence space is essentially infinite, pre-filtering is a necessity, not a novel concept. In this process, practical considerations can not be neglected. We can not reasonably construct any given library – a fact that impacts site selection. Here, oligo pools could be an attractive solution. Their flexibility and low cost could allow for the construction of tailored libraries that were not feasible previously. It should be said, however, that oligo pools are not without flaws. The limited length makes it hard to investigate distant combinations. Additionally, degenerated codons are currently not offered, which means that each specific variant has to be ordered individually, rendering oligo pools uninteresting for combinatorial libraries, even if the mutation sites are close. For example, a five-site combinatorial library would require the purchase of 3.200.000 individual oligonucleotides. The currently very high error rates also limit the applicability of oligos. In our Kemp Eliminase evolution, we estimate error rates of ~ 40 %. At first glance, this seems like a massive drawback of oligo pools. Nevertheless, the incredibly complex libraries constructed at each evolution round should excite any protein engineer. Complete saturation of the entire active pocket, tunnels, as well as distinct mutations spread in the whole gene, can be accomplished with a very manageable physical effort at a reasonable cost. Additionally, oligo pools are a relatively recent product. It is not unreasonable to assume that their quality will improve significantly over the next few years, which would result in a dramatic decrease in the screening burden.

The oligo/filter approach could be valuable to a protein engineer's toolbox. The methodology will not be applicable for every problem, in particular in the very low (<50 samples/round) and high (>10000 samples/round) throughput regime and when dealing with combinatorial effects (where degenerated codons are required or when the mutations are spread further than the maximum oligo length apart). However, the results of this manuscript demonstrate its potential, even in such an early stage of development. With additional automation, reduction in oligo error rates, and increased experience with the overall process, further reducing the required time for a similar HG3 evolution is not unreasonable.

References

[1] S. Lutz and S. M. Iamurri, “Protein engineering: Past, present, and future,” in *Methods in Molecular Biology*, Humana Press Inc., 2018, pp. 1–12. doi: 10.1007/978-1-4939-7366-8_1.

[2] A. Schmid, J. S. Dordick, B. Hauer, A. Kiener, M. Wubbolts, and B. Witholt, “Industrial biocatalysis today and tomorrow,” 2001. [Online]. Available: www.nature.com

[3] T. Nagasawa, M. Wieser, T. Nakamura, H. Iwahara, T. Yoshida, and K. Gekko, “Nitrilase of Rhodococcus rhodochrous J1: Conversion into the active form by subunit association,” *Eur J Biochem*, vol. 267, no. 1, pp. 138–144, 2000, doi: 10.1046/j.1432-1327.2000.00983.x.

[4] E. M. M. Abdelraheem, H. Busch, U. Hanefeld, and F. Tonin, “Biocatalysis explained: From pharmaceutical to bulk chemical production,” *Reaction Chemistry and Engineering*, vol. 4, no. 11. Royal Society of Chemistry, pp. 1878–1894, Nov. 01, 2019. doi: 10.1039/c9re00301k.

[5] R. di Cosimo, J. Mc Auliffe, A. J. Poulose, and G. Bohlmann, “Industrial use of immobilized enzymes,” *Chem Soc Rev*, vol. 42, no. 15, pp. 6437–6474, Jul. 2013, doi: 10.1039/c3cs35506c.

[6] E. L. Bell *et al.*, “Biocatalysis,” *Nature Reviews Methods Primers*, vol. 1, no. 1. Springer Nature, Dec. 01, 2021. doi: 10.1038/s43586-021-00044-z.

[7] A. Glieder, “Protein Engineering Handbook. Edited by Stefan Lutz and Uwe T. Bornscheuer,” *Chembiochem*, vol. 10, pp. 2111–2112, Feb. 2009, doi: 10.1002/cbic.200900410.

[8] R. Blomberg *et al.*, “Precision is essential for efficient catalysis in an evolved Kemp eliminase,” *Nature*, vol. 503, no. 7476, pp. 418–421, 2013, doi: 10.1038/nature12623.

[9] F. Meyer *et al.*, “Modulating Chemoselectivity in a Fe(II)/α-Ketoglutarate-Dependent Dioxygenase for the Oxidative Modification of a Nonproteinogenic Amino Acid,” *ACS Catal*, vol. 11, no. 10, pp. 6261–6269, May 2021, doi: 10.1021/acscatal.1c00678.

[10] M. Eichenberger *et al.*, “Asymmetric Cation-Olefin Monocyclization by Engineered Squalene–Hopene Cyclases,” *Angewandte Chemie - International Edition*, vol. 60, no. 50, pp. 26080–26086, Dec. 2021, doi: 10.1002/anie.202108037.

[11] M. T. Reetz, L. W. Wang, and M. Bocola, “Directed evolution of enantioselective enzymes: Iterative cycles of CASTing for probing protein-sequence space,” *Angewandte Chemie - International Edition*, vol. 45, no. 8, pp. 1236–1241, Feb. 2006, doi: 10.1002/anie.200502746.

[12] M. Voss *et al.*, “Enzyme engineering enables inversion of substrate stereopreference of the halogenase WelO5*,” *ChemCatChem*, Nov. 2022, doi: 10.1002/cctc.202201115.

[13] O. Alvizo *et al.*, “Directed evolution of an ultrastable carbonic anhydrase for highly efficient carbon capture from flue gas,” *Proc Natl Acad Sci U S A*, vol. 111, no. 46, pp. 16436–16441, Nov. 2014, doi: 10.1073/pnas.1411461111.

[14] M. T. Reetz, P. Soni, L. Fernández, Y. Gumulya, and J. D. Carballeira, “Increasing the stability of an enzyme toward hostile organic solvents by directed evolution based on iterative saturation mutagenesis using the B-FIT method,” *Chemical Communications*, vol. 46, no. 45, pp. 8657–8658, Dec. 2010, doi: 10.1039/c0cc02657c.

[15] M. L. Casey *et al.*, ") Rend. 1st," W. Borsche, 1973. [Online]. Available: https://pubs.acs.org/sharingguidelines

[16] D. Röthlisberger *et al.*, "Kemp elimination catalysts by computational enzyme design," *Nature*, vol. 453, no. 7192, pp. 190–195, 2008, doi: 10.1038/nature06879.

[17] H. K. Privett *et al.*, "Iterative approach to computational enzyme design," *Proc Natl Acad Sci U S A*, vol. 109, no. 10, pp. 3790–3795, 2012, doi: 10.1073/pnas.1118082108.

[18] P. Wang, J. Zhang, S. Zhang, D. Lu, and Y. Zhu, "Using High-Throughput Molecular Dynamics Simulation to Enhance the Computational Design of Kemp Elimination Enzymes," *J Chem Inf Model*, 2023, doi: 10.1021/acs.jcim.3c00002.

[19] A. Broom *et al.*, "Ensemble-based enzyme design can recapitulate the effects of laboratory directed evolution in silico," *Nat Commun*, vol. 11, no. 1, Dec. 2020, doi: 10.1038/s41467-020-18619-x.

[20] E. A. Caselle *et al.*, "Kemp Eliminases of the AlleyCat Family Possess High Substrate Promiscuity," *ChemCatChem*, vol. 11, no. 5, pp. 1425–1430, 2019, doi: 10.1002/cctc.201801994.

[21] F. H. Arnold, "Nobel Lecture: Innovation by Evolution: Bringing New Chemistry to Life," 2018.

[22] S. Lutz, "Beyond directed evolution-semi-rational protein engineering and design," *Current Opinion in Biotechnology*, vol. 21, no. 6. pp. 734–743, Dec. 2010. doi: 10.1016/j.copbio.2010.08.011.

[23] Y. Wang, P. Xue, M. Cao, T. Yu, S. T. Lane, and H. Zhao, "Directed Evolution: Methodologies and Applications," *Chem Rev*, vol. 121, no. 20, pp. 12384–12444, Oct. 2021, doi: 10.1021/acs.chemrev.1c00260.

[24] D. Patsch, M. Eichenberger, M. Voss, U. T. Bornscheuer and R. Buller, "LibGENiE – A bioinformatic pipeline for the design of information-enriched enzyme libraries," *Submitted to Comput. Struct. Biotechnol. J., 2023.*

[25] J. D. Bloom, S. T. Labthavikul, C. R. Otey, and F. H. Arnold, "Protein stability promotes evolvability," *Proceedings of the National Academy of Sciences*, vol. 103, no. 15, pp. 5869–5874, 2006, doi: 10.1073/pnas.0510098103.

[26] O. Alvizo *et al.*, "Directed evolution of an ultrastable carbonic anhydrase for highly efficient carbon capture from flue gas," *Proc Natl Acad Sci U S A*, vol. 111, no. 46, pp. 16436–16441, Apr. 2014, doi: 10.1073/pnas.1411461111.

[27] S. J. Novick *et al.*, "Engineering an amine transaminase for the efficient production of a chiral sacubitril precursor," *ACS Catal*, vol. 11, no. 6, pp. 3762–3770, Apr. 2021, doi: 10.1021/acscatal.0c05450.

[28] A. Meyerhans, J.-P. Vartanian, and S. Wain-Hobson, "DNA recombination during PCR," *Nucleic Acids Research*, vol. 18, no. 7. p. 1687. [Online]. Available: https://academic.oup.com/nar/article/18/7/1687/1096541

[29] M. S. Judo, A. B. Wedel, and C. Wilson, "Stimulation and suppression of PCR-mediated recombination," *Nucleic Acids Res*, vol. 26, no. 7, pp. 1819–1825, 1998, doi: 10.1093/nar/26.7.1819.

[30] B. P. Kuiper, R. C. Prins, and S. Billerbeck, "Oligo Pools as an Affordable Source of Synthetic DNA for Cost-Effective Library Construction in Protein- and Metabolic Pathway Engineering," *ChemBioChem*, vol. 23, no. 7. John Wiley and Sons Inc, Apr. 2022. doi: 10.1002/cbic.202100507.

[31] A. Goldenzweig *et al.*, "Automated Structure- and Sequence-Based Design of Proteins for High Bacterial Expression and Stability," *Mol Cell*, vol. 63, no. 2, pp. 337–346, Apr. 2016, doi: 10.1016/j.molcel.2016.06.012.

[32] M. Childers and V. Daggett, "Insights from molecular dynamics simulations for protein design," *Mol. Syst. Des. Eng.*, vol. 2, Jan. 2017, doi: 10.1039/C6ME00083E.

[33] A. Yokota, H. Takahashi, T. Takenawa, and M. Arai, "Probing the roles of conserved arginine-44 of Escherichia coli dihydrofolate reductase in its function and stability by systematic sequence perturbation analysis," *Biochem Biophys Res Commun*, vol. 391, no. 4, pp. 1703–1707, Apr. 2010, doi: 10.1016/j.bbrc.2009.12.134.

[34] R. S. Fredricksen and C. Swenson, "Relationship between Stability and Function for Isolated Domains of Troponin C," *Biochemistry*, vol. 35, pp. 14012–14026, Dec. 1996, doi: 10.1021/bi961270q.

[35] J.-B. Jomain *et al.*, "Structural and thermodynamic bases for the design of pure prolactin receptor antagonists: X-ray structure of Del1-9-G129R-hPRL," *J Biol Chem*, vol. 282, pp. 33118–33131, Dec. 2007, doi: 10.1074/jbc.M704364200.

[36] M. Torrado, J. Revuelta, C. González, F. Corzana, A. Bastida, and J. Asensio, "Role of Conserved Salt Bridges in Homeodomain Stability and DNA Binding," *J Biol Chem*, vol. 284, pp. 23765–23779, Jul. 2009, doi: 10.1074/jbc.M109.012054.

[37] M. Zakrzewska, D. Krowarsch, A. Wiedlocha, S. Olsnes, and J. Otlewski, "Highly stable mutants of human fibroblast growth factor-1 exhibit prolonged biological action," *J Mol Biol*, vol. 352, no. 4, pp. 860–875, Apr. 2005, doi: 10.1016/j.jmb.2005.07.066.

Supporting Information

Draft: Application of the filter/oligo methodology for the re-evolution of the Kemp eliminase HG3.

David Patsch[a,b], Moritz Voss[a], Uwe T. Bornscheuer[b] and Rebecca M. Buller[a, *]

[a] Zurich University of Applied Sciences, School of Life Sciences and Facility Management, Institute of Chemistry and Biotechnology, Einsiedlerstrasse 31, 8820 Wädenswil, Switzerland

[b] Institute of Biochemistry, Department of Biotechnology & Enzyme Catalysis, Greifswald University, Felix-Hausdorff-Strasse 4, D17487 Greifswald, Germany

Material and Methods

Plasmid constructs

The HG3 and HG3.17 genes were ordered as cloned plasmids in pET28b vector between the restriction sites NcoI/XhoI from Twist Bioscience (USA), including a C-terminal his-tag and stop codon. The corresponding pelB-HG3 and pelB-HG3.17 variants used for screening were obtained by cloning the genes in the pET22b plasmid with the NcoI/XhoI restriction enzymes. The resulting variants contain an N-terminal pelB signal peptide and a C-terminal his-tag. The nucleotide sequences of the genes are given in the supporting information.

General cloning protocol of the oligo libraries

The oligo pool was received from Twist Bioscience in dry form and resuspended in MilliQ-water to yield a DNA concentration of 5 ng uL^{-1}.

In the first step, the oligo pool was amplified by PCR using KAPA HiFi HotStart DNA Polymerase (Roche Molecular Systems, Inc.). The PCR mixture (25 µL) comprised of 5 µL 5X KAPA HiFi Buffer, 0.75 µL deoxynucleoside triphosphates (10 mM each), 0.75 µL forward primer (10 µM; Microsynth), 0.75 µL reverse primer (10 µM; Microsynth), 2 µL oligo pool (2 ng; Twist Bioscience), 0.5 µL KAPA HiFi HotStart DNA Polymerase (0.5 U; Roche Molecular Systems, Inc.), and 15.25 µL MilliQ-water. The reaction was performed as follows: (a) 95 °C for 3 min, (b) 20 cycles at 98 °C for 20 s, 55 °C for 15 s, and 72 °C for 25 s, and (c) 72 °C for 1 min. The PCR product was purified using NucleoSpin Gel and PCR Clean-up kit (Macherey-Nagel) according to the manufacturer’s protocol but with elution in 15 µL MilliQ-water.

Secondly, the individual oligo libraries were separately amplified from the oligo pools with their specific primers (designed using the online tool described in **Article III**). The PCR mixture (25 µL) comprised of 5 µL 5X KAPA HiFi Buffer, 0.75 µL deoxynucleoside triphosphates (10 mM each), 0.75 µL forward primer (10 µM; Microsynth), 0.75 µL reverse primer (10 µM; Microsynth), 1 µL purified oligo pool amplification product (around 50 ng), 0.5 µL KAPA HiFi HotStart DNA Polymerase (0.5 U; Roche Molecular Systems, Inc.), and 16.25 µL MilliQ-water. The reaction was performed as follows: (a) 95 °C for 3 min, (b) 20 cycles at 98 °C for 20 s, 55 °C for 15 s, and 72 °C for 25 s, and (c) 72 °C for 1 min. The PCR product was purified by agarose-gel electrophorese (2% agarose gel) and extracted using NucleoSpin Gel and PCR Clean-up kit (Macherey-Nagel) according to the manufacturer’s protocol but with elution in 15 µL MilliQ-water. In addition to the separate amplification of the oligo libraries from the oligo pool, the flanking region of the genes was amplified analog but with the parent variant as PCR template and designed using the online tool described in **Article III**.

Synthesis of 5-nitrobenzisoxazole

The synthesis of 5-nitrobenzisoxazole was performed analog to the reported protocol by Hollfelder *et al.* [1]. 1,2-Benzisoxazole (5 mL, 5.87 g) was added at 0 °C to concentrated H_2SO_4 (20 mL) until the solution turned yellow. A mixture of concentrated HNO_3 (3.4 mL) and concentrated H_2SO_4 (1.3 mL) was added at 0 °C slowly, and the solution was stirred for 30 min. The reaction product was poured onto an ice/water mixture (1:1, 100 mL), and the formed crystals were collected by filtration, washed with ice-cooled water, and dried. The crude product was purified by normal phase flash chromatography (RediSep column) with cyclohexane and ethyl acetate as mobile phases. The solvent was removed *in vacuo* to yield 5-nitrobenzisoxazole (3.05 g) as colorless needles. ^1H NMR (500 MHz, $CDCl_3$) δ 8.90 (d, 1 H), 8.73 (d, 1 H), 8.51 (dd, 1 H), 7.76 (d, 1 H); ^{13}C NMR (500 MHz, $CDCl_3$) δ 164.4, 147.1, 144.8, 125.6, 121.9, 119.2, 110.5.

Screening of the oligo libraries

Preculture in LB medium (100 µL or 1 mL in MTP or DWB, respectively). Inoculation of 1 mL autoinduction medium and growth o/n at 30 °C, followed by 18 °C o/n.

Assay was performed in 200 µL volume in MTP format, and the culture was diluted in reaction buffer according to their activity.

Expression and purification of the HG3 variants

For the expression of the HG3 variants, the constructs in the pET28b plasmid (without pelB-leader sequence) were transformed in *E. coli* BL21(DE3) cells and incubated as preculture overnight at 37 °C, 140 rpm (5 cm throw) in a 5 mL LB-medium (Lysogeny Broth) supplemented with 50 µg mL^{-1} kanamycin. Five milliliters of the preculture were used for the inoculation of 500 mL LB-medium, supplemented with 50 µg mL^{-1} kanamycin, and incubated at 37 °C, 120 rpm (5 cm throw). The expression of the HG3 protein was induced at an optical density of 0.6 at 600 nm by decreasing the temperature to 18 °C and adding 250 µM IPTG 20 min later. The induced culture was incubated overnight at 18°C, 120 rpm. The cells were harvested by centrifugation for 60 min at 3,700 *g*, 4 °C, and the resulting cell pellet was stored at –20 °C.

For purification of the HG3 protein, the cell pellet was resuspended in 15 mL sonication buffer (500 mM NaCl in 50 mM Tris-HCl, pH 7.4), lysed by ultrasonic treatment (Am: 50%, pulse: 1 sec / 1 sec, time: 1 min; Bandelin, Sonoplus), clarified by centrifugation (1 h at 21,000 *g*, 4 °C), and filtrated using a sterile syringe filter (0.45 µm pore size). The lysate was purified using immobilized metal affinity chromatography at an ÄKTA Pure FPLC system (GE Healthcare) using a 5 mL HisTrap FF crude column (Cytiva Sweden AG) with 500 mM NaCl, 20 mM imidazole in 50 mM Tris-HCl, pH 7.4 and eluted with 500 mM NaCl, 300 mM imidazole in 50 mM Tris-HCl, pH 7.4. Subsequently, the samples were concentrated to >5 mL with 10,000 MWCO centrifugal filters (Amicon Ultra – 15; Merck Millipore Ltd.) desalted with three 5 mL HiTrap Desalting columns (Cytiva Sweden AG) in 20 mM NaCl in sodium phosphate buffer pH 6.0. The samples were aliquoted, frozen in liquid nitrogen, and stored at -80 °C.

Melting temperature determination

The melting temperature of the purified variants was determined with the thermal shift assay by performing an HRM analysis (high-resolution melt) with 2.5 µM purified protein in 100 mM NaCl, 50 mM sodium phosphate buffer pH 7.0 supplemented with 20x SYPRO Orange (protein gel stain, Sigma Aldrich). The melting analysis was performed from 25 - 95 °C and recorded at Ex/Em 470/610 nm (Rotor-Gene-Q, Qiagen). The derived melting profile was analyzed for the inflection point (minimum of the first derivative) to determine the melting temperature (Tm).

Michaelis-Menten characterization

The Michaelis-Menten kinetics of the purified HG3 variants were determined photometrically in 200 uL MTP format on a plate reader (Tecan Sparks). The assay was performed in a concentration range of 50 µM – 2 mM of 5-nitrobenzisoxazole with 10% methanol, 100 mM NaCl, and 50 mM sodium phosphate buffer pH 7, and the enzymes were diluted according to their activity. The product formation was followed at 380 nm (ε_{380nm} 8,835 M^{-1} $MTPpathlength^{-1}$), and the linear initial reaction rates for each substrate concentration were fitted towards the Michealis-Menten equation in GraphPad Prism (Version 9.3.1).

Library design based on evolutionary information

The initial MSA was created with the online tool of HHblits (https://toolkit.tuebingen.mpg.de/tools/hhblits), using the UniRef30_2022_02 database and default parameters. The MSA was further processed with HHfilter (https://toolkit.tuebingen.mpg.de/tools/hhfilter) with the settings: max ident: 90, min seq ident: 30, rest default. Variants were selected based on the consensus and frequency strategy outlined in the HotSpot Wizard overview [2], [3]. This led to 76 additional variants that were included in each library.

Library design based on computational mutation scan

ΔΔG predictions were based on a protocol published by the official Rosetta forums: https://www.rosettacommons.org/node/11126. Each mutant was predicted three times, and the lowest energy obtained was compared to the wild-type energies to calculate differences in free energy.

```
def mutate_repack_func4(pose, target_position, mutant, repack_radius, sfxn, ddg_bbnbrs=1,
verbose=False, cartesian=True, max_iter=None):
    import time
    from pyrosetta.rosetta.core.pack.task import operation

    #logger.warning("Interface mode not implemented (should be added!)")

    if cartesian:

sfxn.set_weight(pyrosetta.rosetta.core.scoring.ScoreTypeManager.score_type_from_name('cart_bon
ded'), 0.5)
        #sfxn.set_weight(atom_pair_constraint, 1)#0.5

sfxn.set_weight(pyrosetta.rosetta.core.scoring.ScoreTypeManager.score_type_from_name('pro_clos
e'), 0)

#logger.warning(pyrosetta.rosetta.basic.options.get_boolean_option('ex1'))#set_boolean_option(
'-ex1', True )
        #pyrosetta.rosetta.basic.options.set_boolean_option( 'ex2', True )

    #Cloning of the pose including all settings
    working_pose = pose.clone()

    #Select mutant residue
    mutant_selector =
pyrosetta.rosetta.core.select.residue_selector.ResidueIndexSelector(target_position)

    #Select all except mutant
    all_nand_mutant_selector =
pyrosetta.rosetta.core.select.residue_selector.NotResidueSelector()
    all_nand_mutant_selector.set_residue_selector(mutant_selector)

    #Select neighbors with mutant
    nbr_or_mutant_selector =
pyrosetta.rosetta.core.select.residue_selector.NeighborhoodResidueSelector()
    nbr_or_mutant_selector.set_focus(str(target_position))
    nbr_or_mutant_selector.set_distance(repack_radius)
    nbr_or_mutant_selector.set_include_focus_in_subset(True)

    #Select mutant and it's sequence neighbors
    seq_nbr_or_mutant_selector =
pyrosetta.rosetta.core.select.residue_selector.PrimarySequenceNeighborhoodSelector(ddg_bbnbrs,
ddg_bbnbrs, mutant_selector, False)

    #Select mutant, it's seq neighbors and it's surrounding neighbors
    seq_nbr_or_nbr_or_mutant_selector =
pyrosetta.rosetta.core.select.residue_selector.OrResidueSelector()
    seq_nbr_or_nbr_or_mutant_selector.add_residue_selector(seq_nbr_or_mutant_selector)
    seq_nbr_or_nbr_or_mutant_selector.add_residue_selector(nbr_or_mutant_selector)

    if verbose:
        print(f'mutant_selector:
{pyrosetta.rosetta.core.select.residue_selector.selection_positions(mutant_selector.apply(work
ing_pose))}')
        print(f'all_nand_mutant_selector:
{pyrosetta.rosetta.core.select.residue_selector.selection_positions(all_nand_mutant_selector.a
pply(working_pose))}')
        print(f'nbr_or_mutant_selector:
{pyrosetta.rosetta.core.select.residue_selector.selection_positions(nbr_or_mutant_selector.app
ly(working_pose))}')
        print(f'seq_nbr_or_mutant_selector:
{pyrosetta.rosetta.core.select.residue_selector.selection_positions(seq_nbr_or_mutant_selector
.apply(working_pose))}')
        print(f'seq_nbr_or_nbr_or_mutant_selector:
{pyrosetta.rosetta.core.select.residue_selector.selection_positions(seq_nbr_or_nbr_or_mutant_s
elector.apply(working_pose))}')


    #Mutate residue and pack rotamers before relax
    #if list(pose.sequence())[target_position-1] != mutant:
        #generate packer task
    tf = TaskFactory()
    tf.push_back(operation.InitializeFromCommandline())
    tf.push_back(operation.IncludeCurrent())

    #Set all residues except mutant to false for design and repacking
    prevent_repacking_rlt = operation.PreventRepackingRLT()
    prevent_subset_repacking = operation.OperateOnResidueSubset(prevent_repacking_rlt,
all_nand_mutant_selector, False )
    tf.push_back(prevent_subset_repacking)

    #Assign mutant residue to be designed and repacked
    resfile_comm =
pyrosetta.rosetta.protocols.task_operations.ResfileCommandOperation(mutant_selector, f"PIKAA
{mutant}")
    resfile_comm.set_command(f"PIKAA {mutant}")
    tf.push_back(resfile_comm)

    #Apply packing of rotamers of mutant
    packer = pyrosetta.rosetta.protocols.minimization_packing.PackRotamersMover()
    packer.score_function(sfxn)
    packer.task_factory(tf)
    if verbose:
        logger.warning(tf.create_task_and_apply_taskoperations(working_pose))
    packer.apply(working_pose)

    #allow the movement for bb for the mutant + seq. neighbors, and sc for neigbor in range,
seq. neighbor and mutant
```

```
    movemap = pyrosetta.rosetta.core.select.movemap.MoveMapFactory()
    movemap.all_jumps(False)
    movemap.add_bb_action(pyrosetta.rosetta.core.select.movemap.mm_enable,
seq_nbr_or_mutant_selector)
    movemap.add_chi_action(pyrosetta.rosetta.core.select.movemap.mm_enable,
seq_nbr_or_nbr_or_mutant_selector)

    #for checking if all has been selected correctly
    #if verbose:
    mm  = movemap.create_movemap_from_pose(working_pose)

    logger.info(mm)

    #Generate a TaskFactory
    tf = TaskFactory()
    tf.push_back(operation.InitializeFromCommandline())
    tf.push_back(operation.IncludeCurrent())
    #tf.push_back(operation.NoRepackDisulfides())

    #prevent all residues except selected from design and repacking
    prevent_repacking_rlt = operation.PreventRepackingRLT()
    prevent_subset_repacking = operation.OperateOnResidueSubset(prevent_repacking_rlt,
seq_nbr_or_nbr_or_mutant_selector, True )
    tf.push_back(prevent_subset_repacking)

    # allow selected residues only repacking (=switch off design)
    restrict_repacking_rlt = operation.RestrictToRepackingRLT()
    restrict_subset_repacking = operation.OperateOnResidueSubset(restrict_repacking_rlt ,
seq_nbr_or_nbr_or_mutant_selector, False)
    tf.push_back(restrict_subset_repacking)


    #Perform a FastRelax
    fastrelax = pyrosetta.rosetta.protocols.relax.FastRelax()
    fastrelax.set_scorefxn(sfxn)

    if cartesian:
        fastrelax.cartesian(True)
    if max_iter:
        fastrelax.max_iter(max_iter)

    fastrelax.set_task_factory(tf)
    fastrelax.set_movemap_factory(movemap)
    fastrelax.set_movemap_disables_packing_of_fixed_chi_positions(True)

    if verbose:
        logger.info(tf.create_task_and_apply_taskoperations(working_pose))
    fastrelax.apply(working_pose)
    return working_pose

def cart_ddg(site,res):

    newpose = pose.clone()
    scores = []

    for i in range(3): #ONLY RUNS ONCE!!!!
        scorefxn = create_score_function("ref2015_cart")
        newpose = mutate_repack_func4(newpose,site, res, 6, scorefxn,verbose = False,
cartesian = True)
        news = scorefxn(newpose)
        scores.append(news)
    return scores

all_as = sorted(list(set(pose.sequence())))
sites = np.arange(1,len(pose.sequence()) + 1)

assert len(all_as) == 20

inputs_ = [[site,res] for site in sites for res in all_as]
print(len(inputs_))
cores = 60 # os.cpu_count()
print(cores)

with multiprocessing.Pool(processes=cores) as pool:
    results = pool.starmap(cart_ddg,inputs_)

import pickle
with open('hg34_base.pkl', 'wb') as f:
    pickle.dump(results, f)
```

Library design based on tunnel and ligand analysis

Tunnel analysis was performed with CAVER [4].

Basic parameters:

probe_radius 0.9

shell_radius 4.0

shell_depth 5.0

frame_weighting_coefficient 1.0

frame_clustering_threshold 1.0

The ligand was transferred from the crystal structure of HG3 PDB: 5RGA. Additional variants were selected based on distances to the ligand and tunnels.

Acknowledgments

We thank Katrin Hecht and David Aregger for the chemical synthesis of the 5-nitrobenzisoxazole.

[1] F. Hollfelder, A. J. Kirby, D. S. Tawfik, K. Kikuchi, and D. Hilvert, “Characterization of proton-transfer catalysis by serum albumins,” *J Am Chem Soc*, vol. 122, no. 6, pp. 1022–1029, Feb. 2000, doi: 10.1021/ja993471y.

[2] J. Bendl *et al.*, “HotSpot Wizard 2.0: automated design of site-specific mutations and smart libraries in protein engineering,” *Nucleic Acids Res*, vol. 44, no. 1, pp. W479–W487, Apr. 2016, doi: 10.1093/nar/gkw416.

[3] J. Bendl *et al.*, “HotSpot Wizard 2.0: automated design of site-specific mutations and smart libraries in protein engineering,” *Nucleic Acids Res*, vol. 44, no. 1, pp. W479–W487, Jun. 2016, doi: 10.1093/nar/gkw416.

[4] E. Chovancova *et al.*, “CAVER 3.0: A Tool for the Analysis of Transport Pathways in Dynamic Protein Structures,” *PLoS Comput Biol*, vol. 8, no. 10, pp. e1002708-, Oct. 2012, [Online]. Available: https://doi.org/10.1371/journal.pcbi.1002708

Article V

How to cite: Angew. Chem. Int. Ed. **2021**, 60, 26080–26086
International Edition: doi.org/10.1002/anie.202108037
German Edition: doi.org/10.1002/ange.202108037

Biocatalysis Hot Paper

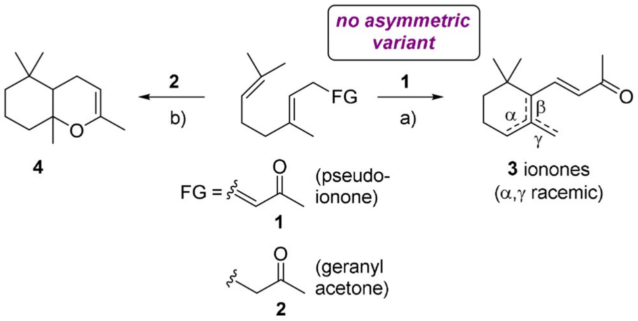
Asymmetric Cation-Olefin Monocyclization by Engineered Squalene–Hopene Cyclases

*Michael Eichenberger[+], Sean Hüppi[+], David Patsch[+], Natalie Aeberli, Raphael Berweger, Sandro Dossenbach, Eric Eichhorn, Felix Flachsmann, Lucas Hortencio, Francis Voirol, Sabine Vollenweider, Uwe T. Bornscheuer, and Rebecca Buller**

***Abstract:** Squalene–hopene cyclases (SHCs) have great potential for the industrial synthesis of enantiopure cyclic terpenoids. A limitation of SHC catalysis has been the enzymes' strict (S)-enantioselectivity at the stereocenter formed after the first cyclization step. To gain enantio-complementary access to valuable monocyclic terpenoids, an SHC-wild-type library including 18 novel homologs was set up. A previously not described SHC (AciSHC) was found to synthesize small amounts of monocyclic (R)-γ-dihydroionone from (E/Z)-geranylacetone. Using enzyme and process optimization, the conversion to the desired product was increased to 79%. Notably, analyzed AciSHC variants could finely differentiate between the geometric geranylacetone isomers: While the (Z)-isomer yielded the desired monocyclic (R)-γ-dihydroionone (>99% ee), the (E)-isomer was converted to the (S,S)-bicyclic ether (>95% ee). Applying the knowledge gained from the observed stereodivergent and enantioselective transformations to an additional SHC-substrate pair, access to the complementary (S)-γ-dihydroionone (>99.9% ee) could be obtained.*

Introduction

Ionones are significant contributors to the appealing scents of many flowers and fruits, including violets, roses, or raspberries.[1] They belong to a family of natural products known as apocarotenoids, which are derived from carotenoids by oxidative cleavage catalyzed by carotenoid oxygenases.[2] An efficient synthetic access to racemic ionones by cation-olefin cyclization of pseudoionone (**1**) was discovered already in the late 19th century by Tiemann and Krüger (Scheme 1).[3] Accordingly, ionones were among the first commercially utilized synthetic fragrance ingredients, featured for example in the iconic fragrance *Vera Violetta* (Roger & Gallet, 1893).



Scheme 1. Cation-olefin cyclizations of pseudoionone (**1**) and geranylacetone (**2**) to racemic ionones (**3**) and bicyclic enolether (**4**), respectively. a) Brønstedt acid b) Brønstedt acid or terpene cyclase.

The organoleptically strongest ionones are the achiral β-ionone and the (*S*)-(+)-isomer of γ-ionone, which has an over 150x lower perception threshold than its optical antipode.[4] Interestingly, this trend is inverted for the corresponding γ-dihydro-analogue, for which the non-natural (*R*)-(−)-isomer has a 6-fold lower perception threshold compared to the (*S*)-(+)-isomer (Scheme 2).[5] The natural (*S*)-(+)-γ-dihydroionone ((*S*)-**5**) occurs for example in Ambergris and is of interest as an intermediate for the synthesis of (−)-α-ambrinol (**6**), which exhibits a highly appreciated animalic scent typical for aged Ambergris tincture (Scheme 2).

The interest in optically active ionones, including both enantiomers of γ-dihydroionone (**5**), has stimulated the development of numerous methods for their synthesis.[5]

---

[*] Dr. M. Eichenberger,[+] S. Hüppi,[+] D. Patsch,[+] Prof. Dr. R. Buller
Zurich University of Applied Sciences, Life Sciences and Facility Management
Einsiedlerstrasse 31, 8820 Wädenswil (Switzerland)
E-mail: rebecca.buller@zhaw.ch

S. Hüppi[+]
Department of Biotechnology, Delft University of Technology
Van der Maasweg 9, 2629 HZ Delft (The Netherlands)

D. Patsch,[+] Prof. Dr. U. T. Bornscheuer
Institute of Biochemistry, Dept. of Biotechnology & Enzyme Catalysis, Greifswald University
Felix-Hausdorff-Strasse 4, 17487 Greifswald (Germany)

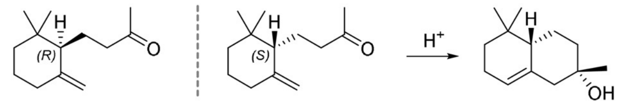
N. Aeberli, R. Berweger, S. Dossenbach, Dr. E. Eichhorn,
Dr. F. Flachsmann, L. Hortencio, F. Voirol
Fragrances S&T, Ingredients Research, Givaudan Schweiz AG
Kemptpark 50, 8310 Kemptthal (Switzerland)

Dr. S. Vollenweider
Science & Technology, Givaudan International SA
Kemptpark 50, 8310 Kemptthal (Switzerland)

[+] These authors contributed equally to this work.

Supporting information and the ORCID identification number(s) for the author(s) of this article can be found under: https://doi.org/10.1002/anie.202108037.

© 2021 The Authors. Published by Wiley-VCH GmbH. This is an open access article under the terms of the Creative Commons Attribution License, which permits use, distribution and reproduction in any medium, provided the original work is properly cited.



(R)-**5** GC-OTH 6 ng/L air (–)-γ-dihydroionone

(S)-**5** GC-OTH 39 ng/L air (+)-γ-dihydroionone

(–)-α-ambrinol (**6**)

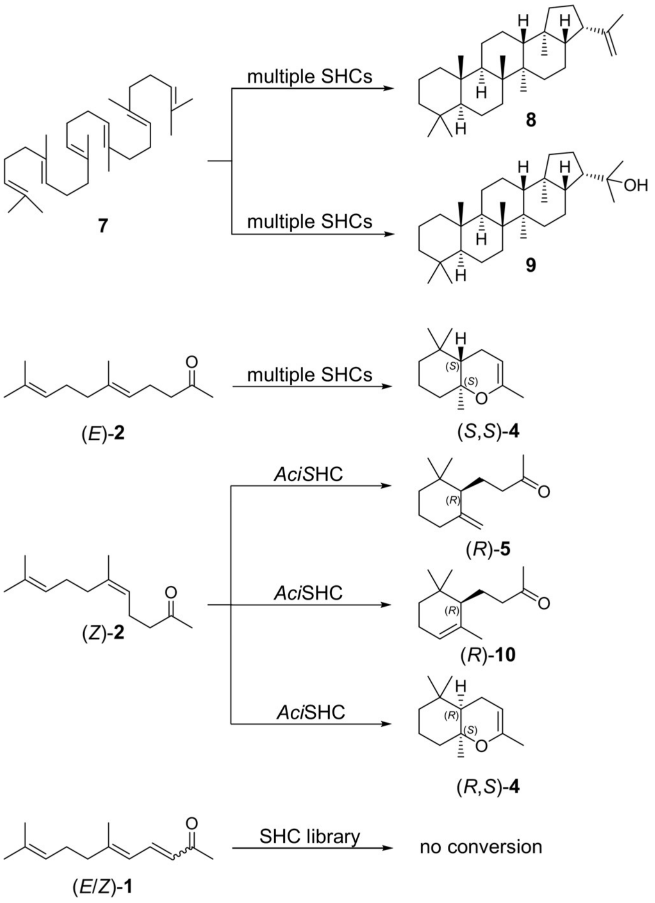
Scheme 2. GC odour thresholds (GC-OTH) of γ-dihydroionone enantiomers (**5**) and conversion of (S)-**5** to (−)-α-ambrinol (**6**).

However, the most obvious synthesis route toward these compounds is currently missing, namely the asymmetric cation-olefin cyclization of pseudoionone (**1**) or a suitable derivative thereof. It appears that carbocation formation at the unpolar isoprene end of the linear chain in combination with enantiospecific folding of the linear C_{13} precursor to form a monocycle is difficult to achieve with classical asymmetric catalysis.

In contrast, squalene–hopene cyclases (SHCs), which belong to the class II terpene cyclases, are capable of locking linear terpenoid substrates in defined chiral conformations, which allows to achieve polyene cyclizations with perfect stereocontrol. Consequently, SHCs have great potential as industrial biocatalysts for the production of enantiopure cyclic terpenoids. A widely spread model reaction is the cyclization of the linear C_{30} triterpene squalene (**7**) into the pentacyclic products hopene (**8**) and hopanol (**9**), through the generation of five new C−C bonds and nine new stereocenters (Scheme 3).[6] The reaction is initiated by the protonation of the unactivated terminal isoprene unit with the unusually acidic middle aspartate of the DXDD active site motif. The excellent chemo-, regio-, and stereocontrol over the polycyclization cascade is achieved through pre-folding of the substrate in a product-like conformation, stabilization and shielding of the highly reactive carbocation intermediates from side reactions, and a selective termination through base assisted proton elimination or addition of water.[7,8] Terpene cyclases from the SHC family are promiscuous enzymes and accept molecules ranging from C_{10} monoterpenoids[9] to C_{35} squalene analogues,[10] and the cyclization reaction can be initiated through protonation of unactivated olefins, carbonyls, and epoxides.[11] This is in contrast to other main families of class II terpene cyclases: oxidosqualene cyclases are limited to substrates containing an epoxide functional group for initial protonation,[12] while class II diterpene cyclases such as *ent*-copalyl diphosphate synthases are generally only active towards the diphosphate containing substrate geranylgeranylpyrophosphate.[13]

Importantly, SHCs have proven to be highly evolvable: Engineered SHC variants with not more than three mutations enabled a viable industrial-scale process to obtain Ambrofix™,[14,15] as well as dramatically increased activity and altered chemo- and stereoselectivity of cyclization reactions with mono- and sesquiterpenoids, such as geraniol,[11] farnesol,[16] or citronellal.[17] A limitation of SHCs, however, is their strict (S)-enantioselectivity at the stereocenter formed after the first cyclization of all polyisoprenoids tested so far (an overview of products is given in reviews[18,19]).

Here, we report our efforts to gain enantio-complementary access to valuable monocyclic terpenoids such as (R)- and (S)-γ-dihydroionone (**5**) via SHC catalysis. Even though the natural diversity of SHC sequences is vast,[20,21] most of the work on non-native substrates has thus far focused on two enzyme variants from *Alicyclobacillus acidocaldarius* (*Aac*SHC) and *Zymomonas mobilis* (*Zmo*SHC1)[19] and only one study reported a screening panel consisting of 12 wild-type enzymes.[22] Thus, to identify enzymes capable of synthesizing (R)- and (S)-γ-dihydroionone (**5**), we opted for a screening approach based on an SHC wild-type library, which included 18 novel SHC homologs. Building on the ability of a newly identified SHC from *Acidothermus cellulolyticus* to generate the monocyclic (R)-γ-dihydroionone ((R)-**5**), we optimized the enzyme by directed evolution and could improve the conversion of nerylacetone ((Z)-**2**) to (R)-γ-dihydroionone ((R)-**5**) by two orders of magnitude to 79 % in 48 h. It should be noted, that during the preparation of this manuscript, a study by the Hauer group was published, which similarly reports the biocatalytic production of (R)-γ-dihydroionone ((R)-**5**) by an engineered SHC from *Alicyclobacillus acidocaldarius.* After five rounds of directed evolution, the authors identified an *Aac*SHC variant with four mutations, which exhibited excellent selectivity (99.5 % *ee*) and conversion (89 %) in seven days.[23]



Scheme 3. Transformations observed in the screening of the SHC library with substrates **1**, **2** and **7**.

In our report, we thus confirm the exciting observation that it is possible to obtain (R)-selective monocyclizations via SHC biocatalysis (> 99 % *ee*) yet using the distinct *Aci*SHC

15213773, 2021, 50, Downloaded from https://onlinelibrary.wiley.com/doi/10.1002/anie.202108037 by Zhaw Zürcher Hochschule Für, Wiley Online Library on [14/06/2023]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

enzyme (51.6% sequence identity to *Aac*SHC). In addition, we observed that all of our *Aci*SHC variants exhibited exquisite selectivity in the transformation of the geometric geranylacetone (**2**) isomers: While the (*Z*)-isomer yielded the desired monocyclic (*R*)-**5** product, the (*E*)-isomer led to the formation of the bicyclic enolether (*S*,*S*)-**4**. Biochemical and docking studies helped us to understand the mechanistic basis of the observed sterodivergent and enantioselective cyclization reactions. Harnessing this knowledge, we ultimately succeeded to additionally obtain the enantio-complementary (*S*)-**5** (> 99.9% *ee*) through the application of an appropriately chosen SHC-substrate pair.

Results and Discussion

In our quest to create an efficient biocatalyst for the enantioselective production of (dihydro-)ionones, we aimed to identify an SHC enzyme with the capability to generate monocyclic products from either (*E*/*Z*)-geranylacetone (**2**) or (*E*/*Z*)-pseudoionone (**1**). *Aac*SHC,[24] *Zmo*SHC1,[24] and engineered variants of these enzymes[25] were previously reported to be inactive towards **1** and were found to convert **2** exclusively into the bicyclic product **4.** Consequently, we chose to explore the SHC diversity beyond these heavily studied variants by setting up a comprehensive screening panel of 31 wild-type enzymes, selected to span all major clades of the phylogenetic tree (Figure S1). The screening library consisted of 13 previously characterized class II terpene cyclases from the SHC-family and 18 novel SHC homologs, which were identified through the presence of two defining PFAM domains for type II triterpene cyclases (PF13249, PF13243) and the SHC-family specific DXDD active site motif (Table S2). As thermostable enzyme scaffolds can be superior starting points for protein engineering and directed evolution approaches,[26] ten of the novel sequences were explicitly chosen to originate from thermophilic bacteria.

To characterize our SHC library and evaluate the biocatalysts' potential for (dihydro)ionone production, we overexpressed the enzymes in *E. coli* BL21(DE3) and carried out whole-cell biotransformations with 10 mM squalene (**7**), 10 mM (*E*/*Z*)-geranylacetone (**2**), and 10 mM (*E*/*Z*)-pseudoionone (**1**). Product formation was analyzed using gas chromatography coupled to mass spectrometry equipped with a flame ionization detector (GC-MS-FID) (Figure 1). Nineteen of the investigated SHCs showed activity towards at least one substrate. Notably, ten of the active enzymes correspond to novel SHC homologs, with sequence identities to experimentally characterized variants between 52.6% and 82.9%. These results validate our bioinformatic search strategy, and the new enzymes further expand the toolbox of available SHCs for biocatalysis.

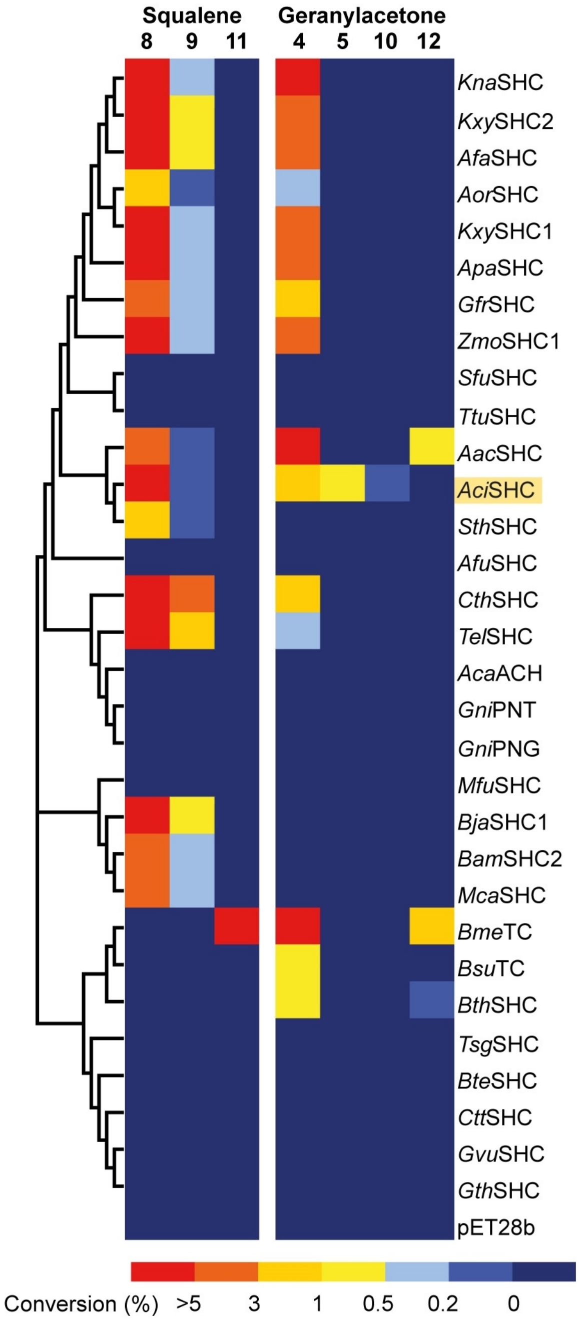


Figure 1. Characterization of the wild-type SHC library with respect to the enzymes' activity towards squalene (**7**) and (*E*/*Z*)-geranylacetone (**2**). Whole-cell biotransformations were carried out by supplementing cell lysate with 10 mM substrate in 50 mM citrate buffer at pH 6 containing 0.8% (**7**) or 0.2% (**2**) of Triton-X-100. The SHCs are ordered based on phylogenetic relationship. Highlighted in yellow is *Aci*SHC, the only wild-type SHC converting **2** into monocyclic products **5** and **10**. Products **11** and **12** could not be structurally assigned.

While (*E*/*Z*)-geranylacetone (**2**) was converted by 15 members of our SHC panel (Figure 1), our screen did not identify any SHC homologs with activity towards pseudoionone (**1**), possibly due to steric and/or electronic effects of the conjugated γ,δ-double bond of **1**, which is the distinguishing feature from **2** (Figure S2). Analyzing the (*E*/*Z*)-geranylacetone (**2**) conversion data in more detail, we identified *Aci*SHC, a novel SHC homolog from the thermophilic bacterium *Acidothermus cellulolyticus*, as a possible candidate for further development. While conversion of the C_{13} substrate **2** into the bicyclic enol ether (**4**) was widespread among the SHC panel, *Aci*SHC was the only enzyme included in the panel that generated two additional minor products

© 2021 The Authors. Published by Wiley-VCH GmbH

15213773, 2021, 50, Downloaded from https://onlinelibrary.wiley.com/doi/10.1002/anie.202108037 by Zhaw Zürcher Hochschule Für, Wiley Online Library on [14/06/2023]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

with conversions of 0.7 % and 0.05 %, respectively. These were identified as γ-dihydroionone (**5**) and α-dihydroionone (**10**) by GC-MS through comparison with authentic reference materials.

Intrigued by these results, we created a sequence alignment of the active pocket[20] of the 14 SHCs, which mainly convert (*E*/*Z*)-geranylacetone (**2**) into the bicyclic product **4** and, in three cases, the structurally unassigned product **12** and compared it to the amino acid distribution of *Aci*SHC (Figure S3). Surprisingly, the sequence alignment revealed that the active site of *Aci*SHC appears to be similar in construction as those of the remaining enzyme panel: Of the 36 residues lining the substrate-binding pocket, only I41, located more than 18 Å away from the catalytic acid D380, was found to be unique in *Aci*SHC (Figure S3). Thus, we proceeded to investigate the unusual product selectivity of *Aci*SHC by constructing its homology model based on the crystal structure of *Aac*SHC (PDB ID: 1SQC; identity: 51.62 %; similarity: 0.45) using SWISS-MODEL[27] followed by docking studies of (*E*)-**2** and (*Z*)-**2** using the software tool AutoDock Vina.[28] Both substrate stereoisomers afforded a docking state with a productive *pre*-chair conformation for monocyclization, however, no "all" *pre*-chair state as required for the formation of the bicyclic product was found (Figure S4).

Thus, even though the identified substrate poses did not fully explain the experimentally observed product distribution, our docking results led us to speculate that already slight changes in the active pocket geometry might result in alternative pre-folding states of **2**. In this way, the enzyme could channel the substrate either into a cationic cascade necessary for the formation of the bicyclic enol ether (**4**) or allow termination of the reaction after a single ring-forming event to yield **5**. In the latter case, deprotonation of the exocyclic methylene group could occur through D378, which in our model of *Aci*SHC is situated at a distance of 2.6 Å from the hydrogen of the relevant carbon C-11. The presence of D378, acting as a catalytic base, could explain the unexpected selectivity for the formation of the energetically unfavorable exocyclic deprotonation product **5** over **10** (Figure S5).

As γ-dihydroionone (**5**) is a compound of particular interest for the flavour and fragrance industry, we aimed to improve the activity and selectivity of the *Aci*SHC catalyzed conversion of **2** into **5** using structure-guided directed evolution. Based on the above-mentioned docking studies of (*E*)-**2** and (*Z*)-**2** into a homology model of *Aci*SHC, we chose 14 sites for NNK single-site saturation libraries. In the first evolution round, we focussed on residues around **2** with the aim to improve pre-folding. In addition, we targeted the large unoccupied space in the active pocket to limit potentially unproductive binding modes known to occur for small substrates in other SHCs (Figure 2 a).[11] Overall, we screened 90 clones for each of the fourteen libraries in deep-well plates amounting to the analysis of >1200 enzyme variants. The screening revealed variants with 2.9 to 5.4-fold increased conversion of **2** into **5** in the libraries A169X, P263X, A310X, G606X, and I613X (Figure 2 b). With the exception of variant G606T, hydrophobic residues were favoured substitutions, and while increased bulk seemed beneficial at sites A169, P263 and A310, smaller amino acids were preferred at I613.

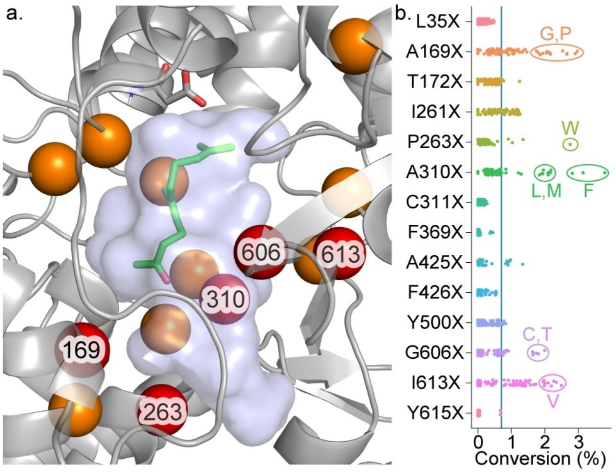


Figure 2. a) Homology model of *Aci*SHC with (*Z*)-**2** docked in the active pocket. Sites colored in orange/red were targeted for single-site saturation mutagenesis, the top-performing sites (red) were then selected for combinatorial mutagenesis in a second evolution round. b) Conversion of (*E*/*Z*)-geranylacetone (**2**) to γ-dihydroionone (**5**) (%FID) by the SHC variants generated in the first round of evolution. The blue line represents the wild-type activity. Top-performing variants are annotated.

Because all beneficial sites were located in the same area of the active pocket of *Aci*SHC, it seemed plausible that epistatic interactions between the amino acid residues might occur. Going forward, we therefore opted to combine all beneficial mutations and the respective wild-type amino acid in a five-site combinatorial library, resulting in a library size of 288 variants. Following library construction by overlap extension PCR, we screened 720 clones for an estimated coverage of 92 %[29] (Figure S6). The best variant for the conversion of (*E*/*Z*)-**2** identified in the second evolution round was dubbed *Aci*SHC_R2.1 (A169P, A310M, G606C, I613V) and achieved a conversion of **2** into **5** of 21.4 %, a 30-fold increase over the wild-type enzyme. In our quest to understand the basis of the increased activity in the engineered *Aci*SHC variants, we sequenced the top ten variants of the second evolution round, all exhibiting a conversion of **2** to **5** of more than 14.4 %. In this analysis, we found nine unique protein sequences with an average of 3.8 mutations. Astonishingly, no single mutation was present in all variants, with the best single site variant, A310F, only occurring in one of the optimized enzymes (Table S3). These findings could indicate that *Aci*SHC can harbor multiple active site geometries, which can induce a productive pre-folding of **2** for efficient cyclization into **5**.

In the analysis of the screening data, we noticed that variants producing high yields of **5** preferentially converted (*Z*)-**2**. To probe this finding further, we carried out whole-cell bioconversions with pure (*E*)-**2** (>99 %) and (*Z*)-**2** (97 %), prepared by fractionated distillation of the mixture of geometric isomers. Using selected enzyme variants spanning the entire evolutionary trajectory, we found that the biotransformations showed intriguing chemoselectivities in the function

15213773, 2021, 50, Downloaded from https://onlinelibrary.wiley.com/doi/10.1002/anie.202108037 by Zhaw Zürcher Hochschule Für, Wiley Online Library on [14/06/2023]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

of the supplied geranylacetone (**2**) geometric isomer: The best three second-round *Aci*SHC variants formed almost exclusively monocyclic products **5** and **10** from (*Z*)-**2** (> 96%) while the bicyclic product **4** was obtained from (*E*)-**2** (> 99%) (Figure 3).

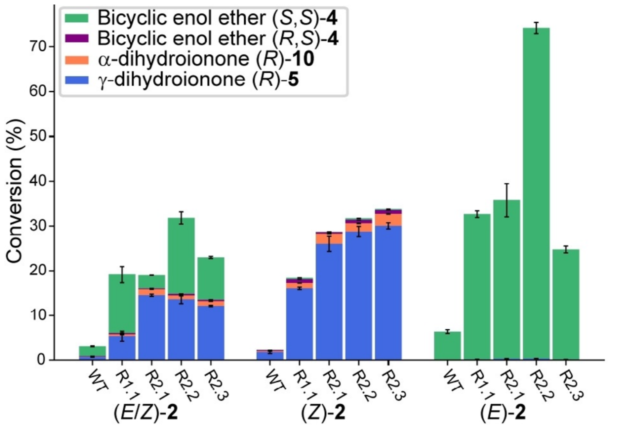
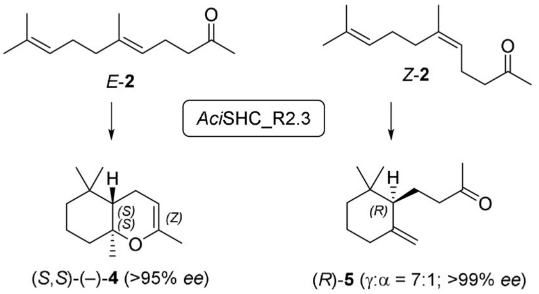


Figure 3. Comparison of the product profile of the wild-type *Aci*SHC with the best variants from each round of enzyme engineering when supplied with 10 mM (*E*/*Z*)-geranylacetone (**2**), nerylacetone ((*Z*)-**2**) or geranylacetone ((*E*)-**2**). The average total recovery was 86 ± 6%.

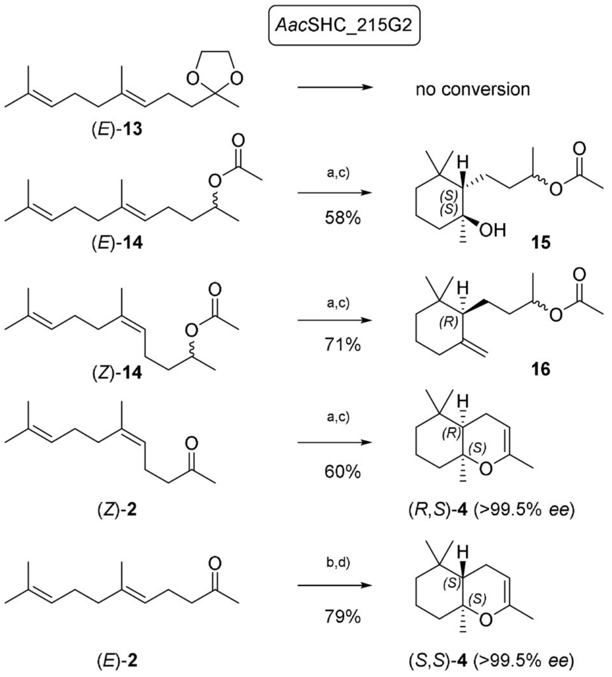
Both products were obtained in excellent optical purity: Using variant *Aci*SHC_R2.3, γ-dihydroionone (**5**) (> 99% *ee*) was formed in the non-natural laevorotatory form, which could be assigned to the absolute (*R*)-configuration based on the work of Brenna et al.,[5] whereas the laevorotatory bicyclic enol ether (**4**) (> 95% *ee*) corresponded to the (*S*,*S*)-configuration as evidenced by comparison to Serra et al.[30] Thus, the SHC enzymes produced the two products **4** and **5** in opposite enantiomeric forms, a process which can be described as a stereodivergent and enantioselective conversion of the (*E*)- and (*Z*)-isomers of **2**. Even when a mixture of (*E*/*Z*)-**2** was used as substrate, **5** was produced as the (*R*)-enantiomer and **4** as the (*S*,*S*)-enantiomer with near to perfect enantioselectivity with all tested variants (Scheme 4, Table S4).

Going forward, we optimized the reaction conditions for *Aci*SHC_R2.3 (A169P, P263W, A310L, I613V), the best second-round variant for the conversion of the geometric isomer (*Z*)-**2** to produce (*R*)-**5** (Figure 3). By optimizing enzyme load (OD 120), temperature (40 °C) and reaction time (48 h), we obtained conversion yields of 79% for the biocatalytic synthesis of (*R*)-**5**, underlining the potential of our engineered SHC variant for manufacturing purposes (Figure S7).



Scheme 4. *Aci*SHC_R2.3 catalyzed stereodivergent and enantiospecific cyclization of (*E*)- and (*Z*)-**2** to (*S*,*S*)-**4** and (*R*)-**5**, respectively.

Going forward, we targeted to evaluate the broader synthetic implications of the observed stereodivergent transformation of geometric isomers by the *Aci*SHC variants. Thus, we set out to transfer our insights to an additional enzyme with the goal to synthesize the (*S*)-enantiomer of γ-dihydroionone ((*S*)-**5**), a key intermediate in the synthesis of (−)-α-ambrinol (**6**).[31] Building on our previous results, we hypothesized that for the synthesis of the natural (*S*)-enantiomer of γ-dihydroionone ((*S*)-**5**), we would require a suitable geranylacetone ((*E*)-**2**) substrate with a masked carbonyl group to prevent the formation of the bicyclic enolether **4**. To that end, the industrially-proven *Aac*SHC variant *Aac*SHC_215G2 was employed for substrate screening in whole-cell biotransformations. Whereas no conversion was observed with dioxolane (*E*)-**13**, we detected the formation of a monocyclic product with intact acetate group from (*E*)-**14** (Scheme 5). To our surprise, the product was not the expected exo-methylene derivative **16**, but its hydrated derivative **15**, formed with perfect enantio- and diastereocontrol. Intrigued by this observation, we repeated the biotransformation with *Aac*SHC_215G2 and (*Z*)-**14**, yielding, as expected, the γ-dihydroionone derivative **16**, again with opposite absolute configuration compared to **15**. These observations prove that the sense of asymmetric induction is



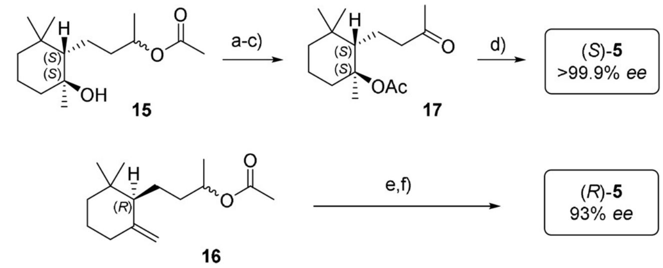
Scheme 5. SHC transformations of (*E*)- and (*Z*)-**14** and **2** with *Aac*SHC_215G2. a) SHC biocatalyst 250 g/l cells, substrate 1.5 g/l (500 mg scale) b) SHC biocatalyst 250 g/l cells, substrate 7 g/l (2 g scale) c) isolated yields after column chromatography d) GC-yield (isolated yields lower due to partial decomposition of products on SiO_2).

© 2021 The Authors. Published by Wiley-VCH GmbH

15213773, 2021, 50, Downloaded from https://onlinelibrary.wiley.com/doi/10.1002/anie.202108037 by Zhaw Zürcher Hochschule Für, Wiley Online Library on [14/06/2023]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

determined solely by the geometry of the double bond in the substrate and is not influenced by the presence of the racemic acetate-bearing chiral center. It is also worth mentioning that no deacetylated product was observed despite the use of a whole cell biocatalyst, where hydrolase-mediated ester hydrolysis could have been expected.

To further explore the scope of different SHC/substrate combinations, we performed whole-cell biotransformations of pure (*E*)- and (*Z*)-**2** with *Aac*SHC_215G2. To our surprise and complementary to the earlier described SHC variants, *Aac*SHC_215G2 converted (*Z*)-**2** to (*R*,*S*)-**4** with perfect enantioselectivity, demonstrating that SHCs can fold a (*Z*)-substrate in such a manner as to form a *cis*-fused bicycle, in line with the *Stork-Eschenmoser* hypothesis.[32] Finally, (*E*)-**2** was converted to (*S*,*S*)-**4** with perfect enantioselectivity and high yield on gram scale by *Aac*SHC_215G2. The chemical transformation of **15** to (*S*)-**5** (Scheme 6) proved the absolute configuration of **15** and provided the first access to the natural (+)-enantiomer of γ-dihydroionone (*S*)-**5** via asymmetric cation-olefin cyclization. The same optical purity of (*S*)-**5** was obtained when tangerinol (**14**; *E*/*Z* 3:2) of commercial quality was used. Similarly, **16** was transformed in two steps to (*R*)-**5**.[33]



Scheme 6. Synthesis of (*S*)- and (*R*)-**5** from cyclotangerinols **15** and **16**. Isolated yields after purification are given. a) acetyl chloride, *N*,*N*-diethylaniline, chloroform, 82%; b) MeOH, $Mg(MeO)_2$, >95%; c) PCC, 54%; d) $NaHCO_3$, DMSO, 140°C, 33% (γ:α:β = 80:9:11); e) K_2CO_3, MeOH, 97% f) PCC, 61%.

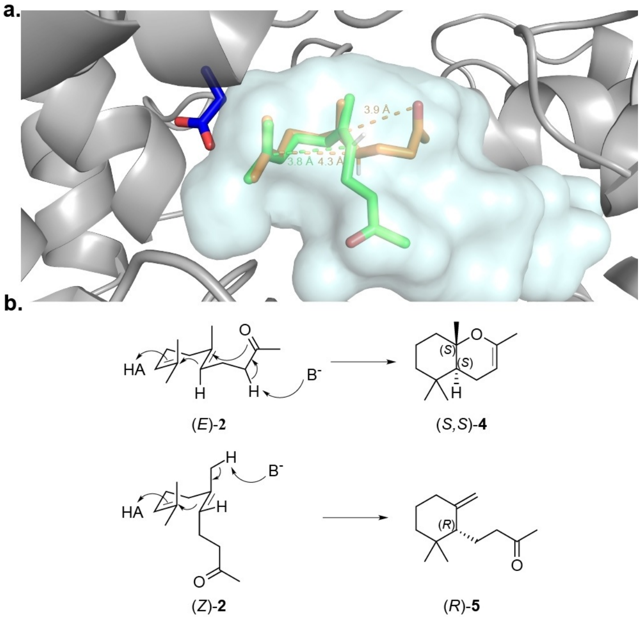


Figure 4. a) Docking of geranylacetone ((*E*)-**2**) (orange) and nerylacetone ((*Z*)-**2**) (green) via AutoDock Vina into a homology model of variant AciSHC_R2.3 prepared by SWISS-MODEL. The corresponding distances for the cyclization reaction are shown. The catalytic aspartate D380 is shown in blue. b) Reaction mechanism for the cyclization of (*E*)-**2** to the bicyclic product (*S*,*S*)-**4** and (*Z*)-**2** to the monocyclic product (*R*)-**5**.

To better understand the mechanistic basis of these stereodivergent and enantioselective reactions, we generated homology models of the engineered *Aci*SHC_R2.3 and *Aac*SHC_215G2 variants using SWISS-MODEL[27] followed by molecular docking of (*Z*)-**2** and (*E*)-**2** as well as (*Z*)-**14** and (*E*)-**14**, respectively, using Autodock Vina.[28] In the homology model of *Aci*SHC_R2.3, (*E*)-**2** showed a reactive all *pre*-chair conformer for generation of a bicyclic product, while for (*Z*)-**2** the second chair was unfolded and the carbonyl-group too distant for an intramolecular nucleophilic attack (Figure 4a). Accordingly, the polycyclization cascade is expected to be interrupted by deprotonation, leading to monocyclic products **5** or **10** (Figure 4b). The pre-folding of the initial chair for (*Z*)-**2** and (*E*)-**2** was nearly identical. Accordingly, the absolute configuration of the newly generated stereocenter resulting from the first cyclization is expected to be defined by the configuration of the double bond (Figure 4). Reflecting our findings for *Aci*SHC, the docking study on *Aac*SHC_215G2 revealed a nearly identical prefolding of the initial *pre*-chairs for (*Z*)-**14** and (*E*)-**14**, suggesting that the enantioselectivity of the cyclization reaction is again determined by the configuration of the double bond of the substrate (Figure S8).

Conclusion

By screening a comprehensive SHC enzyme library, which expands the current SHC toolbox by ten active enzymes, we identified the novel *Aci*SHC capable to cyclize nerylacetone ((*Z*)-**2**) into the monocyclic (*R*)-γ-dihydroionone ((*R*)-**5**). To the best of our knowledge, the recent study by the Hauer group[23] and this work are the first examples of SHCs accepting oxygenated isoprenoids with a (*Z*)-configurated internal double bond, as well as affording a (4a*R*)-stereocenter after the first cyclization. Interestingly, both studies identified similar hotspots in the enzyme active site influencing the monocyclization reaction, albeit in two different enzyme scaffolds with only 51.6% identity (Figure S10 and S11). Notably, through our combinatorial enzyme engineering approach, we identified several highly active *Aci*SHC variants comprising divergent active site geometries, which can afford the necessary pre-folding of **2** to obtain monocyclization products. Our findings therefore indicate that, depending on the enzyme starting scaffold, cyclization cascades cannot only be controlled through the introduction of anchoring hydrogen bonds as shown by Hauer et al.[23] but also through the appropriate choice of the geometric substrate isomer. Transferring this knowledge to the industrially applied *Aac*SHC_215G2 variant, we could highlight that stereodiver-

© 2021 The Authors. Published by Wiley-VCH GmbH

15213773, 2021, 50, Downloaded from https://onlinelibrary.wiley.com/doi/10.1002/anie.202108037 by Zhaw Zürcher Hochschule Für, Wiley Online Library on [14/06/2023]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

gent and enantioselective transformations of geometric isomers could indeed prove to be a general principle in SHC catalysis. Through appropriate substrate engineering and downstream processing, we can obtain access to both enantiomers of a target product via SHC biocatalysis, including the industrially highly relevant chiral building block (*S*)-γ-dihydroionone ((*S*)-**5**). Overall, this work provides an exciting opportunity of tuning the absolute configuration of the cyclized products using substrates with defined double bond stereochemistry and highlights the possibility to control the polycyclization cascade through substrate engineering.

Acknowledgements

The work was supported by the Swiss State Secretariat for Education, Research and Innovation (Federal project contributions 2017–2020: P-14 Innovation in Biocatalysis) and was created as part of NCCR Catalysis, a National Centre of Competence in Research funded by the Swiss National Science Foundation. Open access funding provided by Zurcher Hochschule fur Angewandte Wissenschaften.

Conflict of Interest

The authors declare no conflict of interest.

Keywords: chemoenzymatic synthesis · cyclization · protein engineering · squalene–hopene cyclases · substrate engineering

---

[1] C. Zhang, in *Progress in Carotenoid Research*, IntechOpen, London, **2018**, pp. 85–105.
[2] M. H. Liang, J. Zhu, J. G. Jiang, *Crit. Rev. Food Sci. Nutr.* **2018**, *58*, 2314–2333.
[3] F. Tiemann, P. Krüger, *Ber. Dtsch. Chem. Ges.* **1893**, *26*, 2675–2708.
[4] C. Fuganti, S. Serra, A. Zenoni, *Helv. Chim. Acta* **2000**, *83*, 2761–2768.
[5] E. Brenna, C. Fuganti, S. Serra, P. Kraft, *Eur. J. Org. Chem.* **2002**, 967–978.
[6] G. Siedenburg, D. Jendrossek, *Appl. Environ. Microbiol.* **2011**, *77*, 3905–3915.
[7] D. W. Christianson, *Chem. Rev.* **2006**, *106*, 3412–3442.
[8] K. U. Wendt, *Angew. Chem. Int. Ed.* **2005**, *44*, 3966–3971; *Angew. Chem.* **2005**, *117*, 4032–4037.
[9] G. Siedenburg, D. Jendrossek, M. Breuer, B. Juhl, J. Pleiss, M. Seitz, J. Klebensberger, B. Hauer, *Appl. Environ. Microbiol.* **2012**, *78*, 1055–1062.
[10] I. Abe, H. Tanaka, H. Noguchi, *J. Am. Chem. Soc.* **2002**, *124*, 14514–14515.
[11] S. C. Hammer, A. Marjanovic, J. M. Dominicus, B. M. Nestl, B. Hauer, *Nat. Chem. Biol.* **2015**, *11*, 121–126.
[12] D. W. Christianson, *Chem. Rev.* **2017**, *117*, 11570–11648.
[13] R. J. Peters, *Nat. Prod. Rep.* **2010**, *27*, 1521–1530.
[14] E. Eichhorn, E. Locher, S. Guillemer, D. Wahler, L. Fourage, B. Schilling, *Adv. Synth. Catal.* **2018**, *360*, 2339–2351.
[15] N. Armanino, J. Charpentier, F. Flachsmann, A. Goeke, M. Liniger, P. Kraft, *Angew. Chem. Int. Ed.* **2020**, *59*, 16310–16344; *Angew. Chem.* **2020**, *132*, 16450–16487.
[16] L. C. Kühnel, B. M. Nestl, B. Hauer, *ChemBioChem* **2017**, *18*, 2222–2225.
[17] S. A. Bastian, S. C. Hammer, N. Kreß, B. M. Nestl, B. Hauer, *ChemCatChem* **2017**, *9*, 4364–4368.
[18] I. Abe, *Nat. Prod. Rep.* **2007**, *24*, 1311–1331.
[19] P. O. Syrén, S. Henche, A. Eichler, B. M. Nestl, B. Hauer, *Curr. Opin. Struct. Biol.* **2016**, *41*, 73–82.
[20] S. Racolta, P. B. Juhl, D. Sirim, J. Pleiss, *Proteins Struct. Funct. Bioinf.* **2012**, *80*, 2009–2019.
[21] T. Frickey, E. Kannenberg, *Environ. Microbiol.* **2009**, *11*, 1224–1241.
[22] G. Siedenburg, M. Breuer, D. Jendrossek, *Appl. Microbiol. Biotechnol.* **2013**, *97*, 1571–1580.
[23] A. Schneider, P. Jegl, B. Hauer, *Angew. Chem. Int. Ed.* **2021**, *60*, 13251–13256; *Angew. Chem.* **2021**, *133*, 13359–13365.
[24] M. Seitz, P. O. Syrén, L. Steiner, J. Klebensberger, B. M. Nestl, B. Hauer, *ChemBioChem* **2013**, *14*, 436–439.
[25] S. C. Hammer, *PhD Thesis*, Universität Stuttgart, **2014**.
[26] J. D. Bloom, S. T. Labthavikul, C. R. Otey, F. H. Arnold, *Proc. Natl. Acad. Sci. USA* **2006**, *103*, 5869–5874.
[27] A. Waterhouse, M. Bertoni, S. Bienert, G. Studer, G. Tauriello, R. Gumienny, F. T. Heer, T. A. P. De Beer, C. Rempfer, L. Bordoli, R. Lepore, T. Schwede, *Nucleic Acids Res.* **2018**, *46*, W296–W303.
[28] O. Trott, A. J. Olson, *J. Comput. Chem.* **2009**, *31*, 455–461.
[29] M. T. Reetz, D. Kahakeaw, R. Lohmer, *ChemBioChem* **2008**, *9*, 1797–1804.
[30] S. Serra, V. Lissoni, *Eur. J. Org. Chem.* **2015**, 2226–2234.
[31] P. Naegeli, Y. Wirz-Habersack, *Tetrahedron: Asymmetry* **1992**, *3*, 221–222.
[32] A. Eschenmoser, L. Ruzicka, O. Jeger, D. Arigoni, *Helv. Chim. Acta* **1955**, *38*, 1890–1904.
[33] The lower optical purity of this product can most likely be attributed to a contamination of substrate (*Z*)-**14** with 5 % of the corresponding (*E*)-isomer, which was cyclized to **15**. Partial dehydration of **15** during the bioconversion yields (*S*)-**16**.

Manuscript received: June 17, 2021
Revised manuscript received: July 9, 2021
Accepted manuscript online: August 4, 2021
Version of record online: September 17, 2021

© 2021 The Authors. Published by Wiley-VCH GmbH

15213773, 2021, 50, Downloaded from https://onlinelibrary.wiley.com/doi/10.1002/anie.202108037 by Zhaw Zürcher Hochschule Für, Wiley Online Library on [14/06/2023]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

Angewandte Chemie
Eine Zeitschrift der Gesellschaft Deutscher Chemiker

Supporting Information

Asymmetric Cation-Olefin Monocyclization by Engineered Squalene–Hopene Cyclases

*Michael Eichenberger[+], Sean Hüppi[+], David Patsch[+], Natalie Aeberli, Raphael Berweger, Sandro Dossenbach, Eric Eichhorn, Felix Flachsmann, Lucas Hortencio, Francis Voirol, Sabine Vollenweider, Uwe T. Bornscheuer, and Rebecca Buller**

anie_202108037_sm_miscellaneous_information.pdf

Table of Contents

A. Supplementary Methods 1

B. Supplementary Figures 10

C. Supplementary Tables 21

D. NMR Data 25

E. Chiral GC-analysis of SHC-cyclization products 36

F. Sequences 38

G. References 50

A. Supplementary Methods

1. Materials

Sigma, VWR or Carl Roth were the suppliers for all used chemicals. Molecular genetics were carried out with Phusion polymerase, T4-DNA ligase and restriction enzymes from New England Biolabs. Genes and plasmids were ordered from Twist Bioscience and all oligonucleotides from Microsynth AG.

2. Physical measurements

Analytical GC was performed using an Agilent 8890 GC system with an FID and a single quadrupole MSD (5977B MSD) system (Agilent Technologies, California, USA).

3. Selection of SHC variants

We found that all previously characterized squalene-hopene cyclases (SHCs) contain two defining PFAM domains (PF13249, PF13243). In order to access additional natural diversity of SHCs, we extracted all 5633 protein sequences containing these two domains from Uniprot (accessed on 05.03.2018). We aligned the sequences using ClustalOmega [1] and constructed an approximate maximum likelihood tree using FastTree 2.1.10 [2] with default settings. The tree was divided into two major groups, as previously reported [3] – the oxidosqualene cyclases (OSC) of mainly eukaryotic origin and the SHCs of mainly prokaryotic origin.
We then selected a total of 31 enzymes for our SHC library, chosen to span all major parts of the SHC group of the phylogenetic tree (Table S2). Beyond the heavily studied *Aac*SHC and ZmoSHC1, we included a range of biochemically characterized enzymes, six bacterial SHCs with classical SHC activity on squalene [4–8], as well as three fern [9,10] and two *Bacillus* [11,12] terpene cyclases from the SHC family, catalyzing unusual cyclizations of squalene. We also included a total of 18 not previously characterized homologs. Ten putative SHCs originated from thermophilic bacteria, as well as six SHC homologs from the phylogenetic clade containing ZmoSHC1, in order to explore the sequence space around this enzyme with unusually high activity and promiscuity for small C_{10}-C_{20} substrates [5]. Two final variants were from *Bacillus thuringiensis* B 4219, a strain available in our laboratories and *Aspergillus fumigatus* A1163, which derives from a branch of the phylogenetic tree containing fungal SHCs, which had never been experimentally characterized before.

4. Plasmid construction and enzyme expression in 96-well plates

Genes encoding SHCs in pET28b(+) were purchased from Twist Bioscience. Each plasmid was transformed into *E. coli* BL21(DE3). To obtain single colonies, the cells were plated on LB agar plates containing 50 µg/ml kanamycin. Single colonies were picked into 96 well plates to inoculate 1 ml of LB medium containing 50 µg/ml kanamycin for overnight cultures. From these cultures, a glycerol stock containing 100 µl of culture and 100 µl of 50% Glycerol in water was created. For the expression, 500 µl of Zym-5052 auto-induction medium was inoculated with 50 µl from the overnight culture. Expression was carried out at 20 °C for 24 hours with 300 rpm (5 cm shaking diameter) in a Duetz system for plates. Subsequent, the cells were pelleted by centrifugation (4000g) at 4 °C for 15 minutes. The pellet was washed twice with potassium phosphate buffer (100 mM, pH 7). The cell pellet was immediately used for biotransformation reactions or frozen at – 80°C for later usage.

5. Ligand docking and homology modelling

Homology models of SHCs were created with the webserver SWISS-MODEL[13] and default parameters. The crystal structure of *Aac*SHC (PDB ID: 2SQC) or homology models were used for substrate docking simulations. The ligands were prepared in Chemdraw and energy minimized in Chem3D before saving as mol2 files. The docking was carried out with the AutoDock Vina[14] plugin for UCSF Chimera while using the default parameters. The docking results were visually inspected in Pymol.

6. *Aci*SHC single site saturation libraries, biotransformation, and GC-MS analysis

NNK libraries were created by overlap extension PCR using pSHC8 as a template, with primers listed in Table S1. For each library, two fragements were generated in a first round of PCRs. Specifically, the first fragment was created using the universal forward primer T7_fw and the reverse primer designed for the respective library (e.g. L35X_rv). The second fragment was generated using the library forward primer (e.g. L35X_fw) and the universal reverse primer T7_rv. In all cases, the library forward primer contained the reverse complement of the library reverse primer (overlap), the desired mutation and a silent mutation directly after the desired mutation. All PCRs were carried out with 30 seconds of initial denaturation at 98 °C followed by 29 cycles of 98 °C for 10 seconds, 55°C for 20 seconds and 72 °C for 60 seconds. The final extension was done at 72°C for 10 minutes, and the reactions mixtures were stored at 10 °C until purification. The obtained PCR products were purified by gel extraction, and the DNA concentration was measured via a nanodrop device. Following this step, the two fragments from the first PCRs were used in equimolar amounts as template for a second PCR reaction using primers T7_fw and T7_rv, where they were assembled through the overlap of 15-26bp. The product of the second PCR was purified by gel extraction and afterwards digested with *Xho*I and *Xba*I at 37 °C for 1.5 hours. The digested secondary PCR reactions were ligated into a pET28b(+) vector, which was digested with the same restriction enzymes, using T4 DNA ligase according to the manufacturer's protocols.The DNA sequence of library variants was confirmed by the DNA sequencing service provided by Microsynth AG.

The plasmids were expressed in BL21(DE3) in 96-well plates, as described in 4. The obtained cell pellets were resuspended in half the culture volume of (250 µl) sodium citrate buffer (50 mM, pH 6) with 0.2% Triton-X100 and substrate (10 mM). The screening of native SHCs as well as variants of *Aci*SHC towards (*E*/*Z*)-geranylacetone and (*E*/*Z*)-pseudoionone was carried out in 96-well plates, which were sealed with a heat sealer (Thermo Scientific Alps30) with an alumina seal (Axygen, pierceable sealing film). The seals were chosen to minimize loss of substrates through evaporation. The further characterization of *Aci*SHC variants as well as the screening of native SHCs towards squalene was preformed in glass vials with crimped lids to avoid evaporation. The biotransformations were incubated at 30°C with 300 RPM in the Duetz system for 24 hours. The reactions were stopped with the addition of 800 µl ethyl acetate and a mixing time of 20 minutes at 300 RPM in the Duetz system. Afterwards, they were centrifuged at 4000g, 4 °C for 10 minutes.

Each biotransformation sample was analyzed by GC-MS. The upper layer of ethyl acetate was injected into an Agilent 8890 GC-MS system equipped with a single quadrupole MSD and FID over a J&W DB-5ms GC column (30 m x 0.25 mm x 0.25 µm). Helium was used as the carrier gas. For biotransformations of (*E*/*Z*)-geranylacetone, the GC oven was kept isothermal at 130 °C for 10.5 minutes. The flow was set to be constant at 1.5 ml/min. The MSD was scanning from 30-300 m/z at a speed of 6250. For biotransformations with squalene, the oven was adjusted to 130 °C to 220 °C with 40 °C/min followed by an increase to 310 °C with ten °C/min and the MSD was scanning from 30-500 m/z at a speed of 6250 The MSD was used for the identification of substrate and products. The area detected by the FID was used for the calculation of the conversions (area of product peak divided by the summed area over all product and substrate peaks). The total recovery was calculated using an external calibration curve with authentic geranylacetone reference material.

7. Combination of beneficial mutations

We created a library of *Aci*SHC by recombining the best variants at positions:

A169(A,G,P)
P263(P,W)
A310(A,F,M,L)
G606(G,T,C)
I613(I,V,A,L)

The primers were designed to minimize redundancy (each amino acid is only represented by one codon). Primer sequences can be found in Table S1. The variants were created by overlap extension PCR. In a first round of PCRs, the gene was amplified in five fragments using following primer pairs:

fragment 1: T7_fw, A169_rv,
fragment 2: 1:1:1 mix of A169W_fw:A169G_fw:169P_fw and P263_rv,
fragment 3: 1:1 mix of P263WT_fw:P263W_fw and A310_rv,
fragment 4: 1:1:1:1 mix of A310WT_fw:A310F_fw:A310L_fw:A310M_fw and G606_I613_rv,
fragment 5: 3:1:6:2 mix of G606ACC_I613VTT_fw: G606ACC_I613GCA_fw: G606KGC_I613VTT_fw: G606KGC_I613GCA_fw and T7_rv.

These five fragments were mixed in equimolar amounts for assembly in a second PCR using T7_fw and T7_rv primers and cloned into pET28b(+) as described in 6. The total theoretical library size was 288. Overall, we screened 720 variants for 92% library coverage at 10 mM (*E*/*Z*)-geranylacetone substrate load as described in 6.

8. Preparative scale biotransformation with *Aci*SHC variants

For biotransformation on a preparative scale, cells were cultured overnight at 37 °C at 300 rpm in LB-medium containing 50 µg/ml kanamycin. This pre-culture was then used to inoculate Zymo5052 auto-induction medium in a 1:9 ratio. Protein expression was carried out in baffled shake flasks at 20 °C for 24 hours with 180 rpm. Afterwards, cells were centrifuged (4000g) for 15 minutes and washed over two steps with potassium phosphate buffer (100 mM, pH 7). The cell pellet was resuspended to OD 20 in sodium citrate buffer (50 mM, pH 6) with 0.2% Triton X100 before 40 mM of the substrate was added. The reaction mixture was then incubated at 20 °C for 24 hours at 180 rpm. To increase product concentration, the reaction was quenched and extracted with a much lower ratio of ethyl acetate: reaction volume (1:8) compared to plate screening.

9. Bioconversion reactions with *Aac*SHC variant 215G2

9.1. 215G2 SHC biocatalyst production

The biocatalyst of the *Aac*SHC variant 215G2 was produced in fermentations as described elsewhere.[15]

9.2. Small scale cyclization of tangerinol isomers

Cyclization reactions (1 ml total volume) with 1 g/l *E*- and *Z*- tangerinol (*E*- and *Z*-14) were run with cells that had produced *Aac*SHC 215G2 at OD650nm of 40 and 0.28% SDS in 0.1 M succinic acid/NaOH buffer pH 5.4. The reactions were incubated at 35°C, and under constant agitation (900 rpm, Heidolph Synthesis 1 Liquid 24). After 24h reaction time, the biotransformations were extracted with 1 ml tert-butylmethylether (MTBE) for GC-MS analysis.

9.3. Preparative scale biotransformations with tangerinol and geranylacetone

Bioconversions were run in 350 ml total volume (750 ml InforsHT reactors) with 250 g/l (wet weight) of cells that had produced *Aac*SHC 215G2 in 0.1 M succinic acid/NaOH buffer at pH 5.4, and 35 °C under constant agitation (700 rpm). They contained 1.30% SDS (w/v), 1.5 g/l nerylacetone (*Z*-**2**), 7 g/l geranylacetone (*E*-**2**), 1.5 g/l racemic *E*-tangerinol (*E*-**14**), or 1.5 g/l racemic *Z*-tangerinol (*Z*-**14**). Cell suspensions were prepared by suspending frozen cell pellets in reaction buffer. The cell concentration was determined by centrifuging an aliquot of cell suspension (17210 *g*, 10 min, 4°C), and the required amount of cell suspension calculated for 250 g/l wet weight of cells in the reaction. SDS was added from a 31% SDS stock solution in deionized water. The required volume of succinic acid buffer for a total reaction volume of 350 ml was deduced. To the reactor was added in the following order: substrate, SDS, buffer, cell suspension. The pH of the reactions was carefully set to 5.4 dropwise with 85% H_3PO_4. pH in the reaction was monitored *in situ* and controlled with a calibrated external electrode. pH adjustment was done with 10% H_3PO_4 if required. Conversion was determined by GC-analysis calculating the ratio of product and peak surface areas (details see below). The bioconversion reactions were finally extracted three times with *tert*-butylmethylether (MTBE): 150, 100, and 100 ml. The fractions were analyzed for their product and substrate content by GC-analysis after dilution if required and finally pooled for further processing.

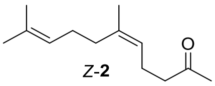
9.4. Sample preparation and gas chromatography analysis

Bioconversion reactions were sampled over time (200 µl samples), and the samples extracted with 200 µl MTBE. After centrifugation (Eppendorf centrifuge 5415C, 14000 rpm), the solvent phase was analyzed by GC-FID. 1 µl solvent phase was injected (split ratio 10, split flow 40 ml/min) onto a 30 m x 0.32 mm x 0.25 µm DB-5 column (Trace 1310 gas chromatograph, Thermo). The column was developed at 4 ml/min H_2 constant flow: 100 °C, 15 °C/min to 200 °C, 120 °C/min to 240 °C, 3 min at 240 °C; inlet and detector temperature: 250 °C.

10. Synthesis of substrates and synthetic transformation of SHC products to γ-dihydroionone

All reagents and reaction solvents were analytical grade, purchased from commercial suppliers and used without further purification. Reactions were monitored by GC-FID (Zebron ZB-5 GC capillary column, 12 m x , 0.32 mm x 0.25 μm, or Zebron ZB-wax, 15 m x , 0.32 mm x 0.25 μm). Flash column chromatography was performed on Biotage silica gel prepacked columns (particle size 20 μm) with the eluent indicated eluents, flow 50 ml/min. All reported yields, unless otherwise specified, refer to spectroscopically and chromatographically pure isolated compounds; isomeric ratios are indicated if appropriate. Routine NMR spectra were recorded on Bruker Avance III HD, 2D NMR spectra were recorded on Bruker Avance-III 600 MHz with 1.7 mm TCI-microcryoprobe). Proton chemical shifts are reported in ppm (δ) relative to tetramethylsilane (TMS), with the solvent resonance employed as the internal standard ($CDCl_3$ δ 7.27 ppm). Data are reported as follows: chemical shift, multiplicity (s = singlet, d = doublet, t = triplet, q = quartet, p = pentet, sext = sextet, h = heptet, m = multiplet, br = broad), coupling constants (Hz) and integration. ^{13}C chemical shifts are reported in ppm from tetramethylsilane (TMS) with the solvent resonance as the internal standard ($CDCl_3$ δ 77.0 ppm;). Mass spectra were recorded with GC-MS (Agilent Technologies 7890A / 5975C) equipped with an SGE BPX5 capillary column (12 m, 0.22 mm i.d. 0.25 μm film) operated at a constant He flow of 1 ml/min. The temperature program started at 50 °C (for 2 min), then with 20 °C/min to 240 °C followed by 35 °C/min to 270 °C (for 3 min). The MS transfer line and ion source temperatures were 250 °C and 230 °C, respectively. The quadrupole MS was equipped with an EI ion source at 70 eV. For high-resolution mass spectra, samples were diluted in methanol and directly introduced in the MS (Thermo Scientific, Q Exactive Orbitrap) by a syringe pump (Chemyx Inc., Fusion 100 T) at a flowrate of 5-10 μl/min. Data was recorded with 70.000 mass resolution using Xcalibur (4.1.50) and analyzed with Xcalibur (4.2.28.14). Optical rotations were determined with Anton Paar MCP 200 Polarimeter at 589 nm and 25 °C. Data are reported as follows: [α]λ temp, concentration (c; g/100 ml), and solvents. Enantiomeric excesses (% e.e.) were determined by GC employing a chiral stationary phase column specified in the individual experiment. The required racemic standards were synthesized in separate experiments (*vide infra*), and the chiral GC methods were optimized to ensure baseline separation of the two enantiomers.

*10.1. Nerylacetone (Z-**2**)*

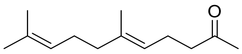


Z-**2**

Commercial geranylacetone (*E*/*Z* 3:2, 767 g) was fractionally distilled over a 1 m Sulzer column (0.08 mbar, bath temperature 125-138 °C). Fractions collected at 82-85 °C contained >92% (*Z*)-isomer (302.5 g). This product (150 g) was further distilled over a 30 cm Sulzer column (0.08 mbar, bath temperature 148 °C), yielding (*Z*)-**2** (100.5 g) as a clear, colourless liquid with an isomeric purity of 97%.

^1H-NMR ($CDCl_3$, 400 MHz): 5.07 - 5.14 (m, 2 H), 2.43 - 2.47 (m, 2 H), 2.23 - 2.30 (m, 2 H), 2.14 (s, 3 H), 2.03 - 2.07 (m, 4 H), 1.68 - 1.70 (m, 6 H), 1.62 (s, 3 H). ^{13}C-NMR ($CDCl_3$, 101 MHz): 208.7 (s), 136.5 (s), 131.6 (s), 124.2 (d), 123.3 (d), 44.0 (t), 31.9 (t), 29.9 (q), 26.5 (t), 25.7 (q), 23.3 (q), 22.3 (t), 17.6 (q). MS (EI, 70eV): 194 (M^+, <1), 151 (39), 136 (15), 125 (7), 107 (12), 93 (11), 69 (56), 43 (100).

*10.2. Geranylacetone (E-**2**)*

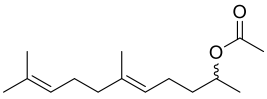


E-**2**

The residue of the above distillation (51 g) contained mainly E-isomer and was distilled over a 10 cm Vigreux column (0.12 mbar, bath temperature 121-122 °C, head temperature 65-73 °C) to yield (*E*)-**2** (35.5 g) as a clear, colourless liquid with an isomeric purity of >99.9%.

^1H-NMR ($CDCl_3$, 400 MHz): 5.03 - 5.12 (m, 2 H), 2.43 - 2.49 (m, 2 H), 2.22 - 2.31 (m, 2 H), 2.13 (br. s, 3 H), 1.94 - 2.10 (m, 4 H), 1.67 (br. s, 3 H), 1.62 (br. s, 3 H), 1.60 (br. s, 3 H). ^{13}C-NMR ($CDCl_3$, 101 MHz): 208.9 (s), 136.4 (s), 131.4 (s), 124.2 (d), 122.5 (d), 43.7 (t), 39.6 (t), 29.9 (q), 26.6 (t), 25.7 (q), 22.5 (t), 17.7 (q), 16.0 (q). MS (EI, 70 eV): 194 (M^+, <1), 151 (11), 136 (10), 125 (6), 107 (11), 93 (6), 69 (35), 43 (100).

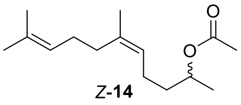
*10.3. rac. E-Tangerinol (E-***14***)*



E-**14**

Sodium borohydride (0.77 g, 20.4 mmol, 0.8 equiv.) was added to the solution of geranylacetone (*E*-**2**) (4.96 g, 25.5 mmol, 1 equiv.) in methanol (80 ml) at 0°C. The mixture was stirred for 3 h at room temperature, then poured on 2N aq. HCl-solution (200 ml). The aqueous layer was extracted with MTBE (2x 100 ml), the combined organic layers were washed with water and brine (pH 6), then dried over $MgSO_4$ and concentrated. Crude rac. (*E*)-6,10-dimethylundeca-5,9-dien-2-ol was obtained as a yellow oil (4.90 g, 98%), which was further acetylated (4.0 g, 20.4 mmol, 1 equiv.) in cyclohexane (50 ml) with acetic anhydride (4.2 g, 40.7 mmol, 2 equiv.) and 1 drop of conc. phosphoric acid at room temperature for 18 h. The mixture was poured on 2N aq. NaOH-solution (50 ml), then worked up as described above to yield 4.93 g of a yellow oil which was purified by flash column chromatography with heptane / MTBE 30:1 to yield (*E*)-**14** (*E*-Tangerinol, 4.45 g, 92%) as a clear, colourless oil.
^{1}H-NMR ($CDCl_3$, 100 MHz): 5.05 - 5.14 (m, 2 H), 4.83 - 4.94 (m, 1 H), 2.03 - 2.10 (m, 2 H), 2.03 (br. s, 3 H), 1.95 - 2.02 (m, 3 H), 1.61 - 1.70 (m, 4 H), 1.60 (br. s, 3 H), 1.59 (br. s, 3 H), 1.45 - 1.55 (m, 1 H), 1.25 - 1.31 (m, 1 H), 1.21 (d, *J*=6.1 Hz, 3 H). ^{13}C-NMR ($CDCl_3$, 101 MHz): 170.7 (s), 135.7 (s), 131.3 (s), 124.2 (d), 123.3 (d), 70.7 (d), 39.7 (t), 35.9 (t), 26.7 (t), 25.7 (q), 23.9 (t), 21.4 (q), 20.0 (q), 17.7 (q), 15.9 (q). MS (EI, 70 eV): 178 ([M-AcOH]$^+$, 8), 163 (6), 135 (21), 123 (10), 109 (100), 93 (17), 81 (21), 69 (78), 43 (72).

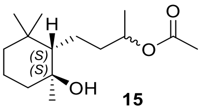
*10.4. rac. Z-Tangerinol (Z-***14***)*



Z-**14**

The procedure described above for the preparation of (*E*)-**14** was repeated with nerylacetone (Z-2) to furnish (*Z*)-Tangerinol (*Z*-**14**, 82% over 2 steps) as a clear, colourless oil.
^{1}H-NMR ($CDCl_3$, 100 MHz): 5.06 - 5.15 (m, 2 H), 4.84 - 4.94 (m, 1 H), 1.96 - 2.09 (m, 6 H), 2.02 (s, 3 H), 1.57 - 1.71 (m, 9 H), 1.46 - 1.54 (m, 1 H), 1.24-1.33 (m, 1 H), 1.21 (d, *J*=6.1 Hz, 3 H). ^{13}C-NMR (CDCl3, 101 MHz): 170.7 (s), 135.8 (s), 131.6 (s), 124.2 (d), 124.2 (d), 70.7 (d), 36.2 (t), 31.9 (t), 26.5 (t), 25.7 (q), 23.8 (t), 23.4 (q), 21.3 (q), 20.0 (q), 17.6 (q). MS (EI, 70 eV): 178 ([M-AcOH]$^+$, 10), 163 (9), 135 (18), 109 (72), 93 (22), 81 (20), 69 (100), 43 (89).

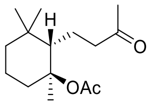
10.5. 4-((1S,2S)-2-hydroxy-2,6,6-trimethylcyclohexyl)butan-2-yl acetate (**15**) *with AacSHC 215G2*



15

The MTBE-extract of the biocatalytic conversion of racemic *E*-Tangerinol (*E*-**14**, 525 mg, 2.2 mmol) with *Aac*SHC 215G2 (250 g/l, 350 ml, cf. chapter 9.3.) was concentrated in a rotary evaporator and the residue was purified via column chromatography (heptane/MTBE 3:2) to yield **15** as a colourless oil (330 mg, 58%, purity by GC/MS > 99.9%, mixture of 2 diastereomers 2:1 according to NMR, not resolved by GC). $[\alpha]_D$ = +2.6 (c = 0.98, $CHCl_3$).
^{1}H-NMR ($CDCl_3$, 100 MHz*,* mixture of 2 diastereomers): 4.84 - 4.95 (m, 1 H), 2.04 (s, 2 H, main isomer), 2.03 (s, 1 H, minor isomer), 1.21-1.89 (m, 10 H), 1.22 (d, J = 6.2 Hz, 1H, minor isomer), 1.22 (d, J = 6.2 Hz, 2H, major isomer), 1.17 (br. s, 1H, minor isomer), 1.16 (br. s, 2H, major isomer), - 1.07-1.13 (m, 1 H), 0.94 (s, 2 H, major isomer), 0.92 (s, 1 H, minor isomer.), 0.81 (s, 3 H). ^{13}C-NMR ($CDCl_3$, 101 MHz): (main isomer) 171.0 (s), 74.1 (s), 71.6 (d), 57.0 (d), 43.7 (t), 41.5 (t), 39.0 (t), 35.5 (s), 32.8 (q), 23.2 (q), 21.9 (t), 21.4 (q), 21.2 (q), 20.5 (t), 20.0 (q). (minor isomer) 170.9 (s), 74.2 (s), 71.5 (d), 56.7 (d), 43.6 (t), 41.5 (t), 38.7 (t), 35.5 (s), 32.7 (q), 23.3 (q), 21.8 (t), 21.4 (q), 21.2 (q), 20.4 (t), 19.9 (q). MS (EI, 70 eV, mixture of diastereomers): 256 (M$^+$, <1), 196 (4), 181 (7), 163 (8), 153 (13), 138 (15), 127 (15), 109 (56), 96 (35), 81 (20), 69 (38), 55 (25), 43 (100). HR-MS (ESI, positive mode): $C_{15}H_{28}O_3$ calcd. for [M+Na]$^+$ 279.1931, found 279.1931.

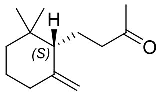
10.6. (1S,2S)-(+)-1,3,3-trimethyl-2-(3-oxobutyl)cyclohexyl acetate (**17**)



17

The solution of **15** (250 mg, 0.98 mmol), acetyl chloride (1.1 g, 15 equiv.) and N,N-diethylaniline (1.41 g, 9.8 mmol, 10 equiv.) in chloroform (20 ml) was refluxed for 3 days unders stirring. The mixture was added to aq. HCl solution (2 M, 30 ml) and extracted MTBE (2x 30 ml). The organic layers were washed with 30 ml of each, 2 M aq. NaOH solution, water and brine. The combined organic layers were dried over $MgSO_4$, filtered and evaporated to yield 4-((1S,2S)-2-acetoxy-2,6,6-trimethylcyclohexyl)butan-2-yl acetate (240 mg, 82%) as a yellow liquid. Selective hydrolysis of the secondary acetate group was effected by stirring the solution of this product in MeOH (20 ml) in the presence of magnesium methanolate (6.8 ml of a 10% solution in MeOH, 8 equiv.) at room temperature for 30 h. Water (20 ml) was added and the mixture was extracted and worked up as described above to yield a yellow liquid (470 mg >100%). The product was dissolved in CH_2Cl_2 (15 ml) and oxidation of the secondary alcohol was effected by stirring with pyridinium chlorochromate (PCC, 553 mg, 2.6 mmol, 1.4 equiv) for 1.5 h. The mixture was filtered over silica gel and the product was purified by flash column chromatography with heptane / MTBE 3:1 to yield (1S,2S)-1,3,3-trimethyl-2-(3-oxobutyl)cyclohexyl acetate (**17**) as a clear, yellow liquid (110 mg, 44% over 3 steps, purity >99.9% according to GC-MS). $[\alpha]_D$ = +24.6 (c = 1.46, $CHCl_3$) ^{1}H-NMR ($CDCl_3$, 400 MHz): 2.42 - 2.67 (m, 3 H), 2.12 (s, 3 H), 1.92 (s, 3 H), 1.51 - 1.68 (m, 5 H), 1.47 (s, 3 H), 1.31 - 1.43 (m, 2 H), 1.16 - 1.25 (m, 1 H), 0.93 (s, 3 H), 0.83 (s, 3 H). ^{13}C-NMR ($CDCl_3$, 101 MHz): 209.0 (s), 169.9 (s), 87.5 (s), 52.8 (d), 46.0 (t), 40.4 (t), 37.3 (t), 35.5 (s), 32.4 (q), 29.8 (q), 22.9 (q), 22.0 (br., q), 20.4 (t), 20.3 (br., q), 19.6 (t). MS (EI, 70 eV): 254 (M^+, <1), 212 ([M-ketene]$^+$, <1), 194 ([M-AcOH]$^+$, 2), 176 (9), 161 (14), 136 (32), 121 (35), 109 (19), 95 (22), 81 (15), 69 (16), 55 (12), 43 (100). HR-MS (ESI, positive mode): $C_{15}H_{26}O_3$ calcd. for [M+Na]$^+$ 277.1774, found 277.1773.

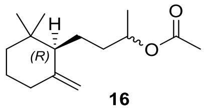
*10.7. (S)-(+)-γ-dihydroionone ((S)-**5**) from* **17**



(*S*)-**5**

To the solution of **17** (60 mg, 0.24 mmol) in DMSO (5 ml) was added $NaHCO_3$ (40 mg, 0.47 mmol, 2 equiv.) and the mixture was heated to 140 °C for 30 h. The mixture was poured into water (30 ml) and extracted with heptane. After standard workup and flash column chromatography (heptane / MTBE 10:1), (*S*)-**5** was isolated as a colourless oil (15 mg, 33%). The product contains 9% (*S*)-α-dihydroionone and 11% β-dihydroionone. NMR and MS spectra were in accordance with published data [*C. Fuganti*, *S. Serra*, *A. Zenoni, Helv. Chim. Acta* **2000**, *83*, 2761.]. Chiral GC-analysis (Astec Chiraldex G-DP): 36.34 min, >99.9% e.e.

10.8. 4-((R)-2,2-dimethyl-6-methylenecyclohexyl)butan-2-yl acetate (**16**) *with AacSHC 215G2*

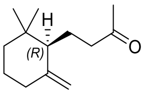


16

The MTBE-extract of a biocatalytic conversion of racemic (*Z*)-**14** (525 mg, 2.2 mmol) with *Aac*SHC 215G2 (250 g/l, 350 ml) as described in chapter 9.3. was concentrated in a rotary evaporator, and the residue was purified via column chromatography (heptane/MTBE 95:5) to yield **16** as a colourless oil (370 mg, 71%). For spectral analysis, a sample was further purified by semipreparative HPLC (normal phase silica, Lichrospher 5 μm, 4 ml/min hexane/ 2.5% MTBE).

^{1}H-NMR (C_6D_6, 600 MHz): 4.99 - 5.06 (m, 1 H), 4.80 - 4.83 (m, 1 H), 4.59 - 4.62 (m, 1 H), 1.91 - 1.95 (m, 2 H), 1.72 (s, 3 H), 1.58 - 1.62 (m, 1 H), 1.37 - 1.51 (m, 5 H), 1.21 - 1.36 (m, 2 H), 1.12 (d, *J*=6.0 Hz, 3 H), 1.06 - 1.11 (m, 1 H), 0.91 (s, 3 H), 0.80 (s, 3 H). ^{13}C-NMR (C_6D_6, 600 MHz, from HSQC): 169. 8 (s), 149.2 (s), 109.6 (t), 71.1 (d), 54.2 (d), 36.0 (t), 35.1 (t), 34.8 (s), 32.3 (t), 28.4 (q), 26.6 (q), 23.9 (t), 22.5 (t), 20.9 (q), 20.0 (q).
MS (EI, 70 eV): 178 ([M-AcOH]$^+$, 5), 136 (55), 122 (29), 109 (58), 93 (61), 81 (31), 69 (49), 55 (25), 43 (100). HR-MS (ESI, positive mode): $C_{15}H_{26}O_2$, calcd. for [M+Na]$^+$ 261.1825, found 261.1825.

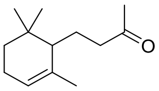
10.9. *Conversion of* **16** *to (R)*-**5**



(*R*)-**5**

The above product (300 mg, 1.26 mmol) was saponified with K_2CO_3 (348 mg, 2.52 mmol, 2 equiv.) in MeOH (50 ml) for 3 h under reflux and stirring. The mixture was cooled to RT, poured on water (40 ml) and extracted with MTBE. The organic layers were washed with water and brine and dried over $MgSO_4$. After filtration and removal of the solvent, a colourless oil was obtained (240 mg), which was dissolved in CH_2Cl_2 (15 ml), pyridinium chlorochromate was added (369 mg, 1.71 mmol, 1.4 equiv.), and the mixture was stirred for 2 h at RT. The mixture was filtered over silica gel, the filter cake was rinsed with MTBE, and the filtrate was dried. The residue was purified by FC (heptane/MTBE 93:7) to yield (*R*)-**5** as a colourless oil (145 mg, 61%). The product contained 11% (R)-(+)-α-dihydroionone. NMR and MS spectra were in accordance with published data.[16] Chiral GC-analysis (Astec Chiraldex G-DP): 36.62 min, 93% e.e.

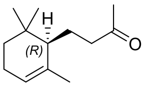
10.10. *Synthesis of rac.-α-dihydroionone (rac.-***10***)*



rac.-**10**

Commercial α-ionone (3.0 g, 15.6 mmol) was hydrogenated in THF (40 ml) in the presence of Raney nickel (400 mg) and H_2 (1 atm) for 5 h. The mixture was filtered by suction, and the filtrate was concentrated i. RV to yield rac. α-dihydroionone (2.94 g, 97%). The product was used for subsequent steps without further purification. A sample was purified by flash chromatography on silica gel as racemic reference for chiral GC-analysis (purity according to GC-MS 91%, NMR data were in accordance with published data,[16] MS data matched with MS-library hit. Chiral GC-analysis (Hydrodex-beta-3P): (S)-(-)-10 (43.55 min), (R)-(+)-10 (44.64 min).

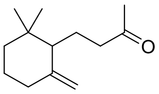
10.11. *Synthesis of (R)-(+)-α-dihydroionone ((R)-***10***)*



(*R*)-**10**

The procedure described in chapter 10.10. was repeated with a commercial sample of (R)-(+)-ionone (200 mg, 1.04 mmol) to yield (*R*)-**10** as a colourless oil (202 mg, >99%, purity according to GC-MS 98%). $[\alpha]_D$ = +136.9 (c = 1.26, EtOH); (Lit. +138.4 (c = 0.65, EtOH),[17]). Chiral GC-analysis (Hydrodex-beta-3P): 44.61 min, >99.9% e.e.

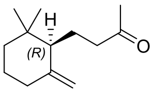
10.12. *Synthesis of rac.γ-dihydroionone (rac.-***5***)*

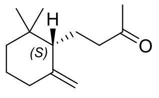


rac.-**5**

To the solution of racemic α-dihydroionone (2.94 g, 15.1 mmol, prepared in chapter 10.10.) in MeOH (40 ml) was added $NaBH_4$ (0.46 g, 12.1 mmol, 0.8 equiv.) at 0°C. The solution was stirred for 15 min; then the cooling bath was removed and stirring continued for 90 min. The solution was poured in 2M aq. HCl-solution (100 ml) and extracted with MTBE. The organic layers were washed with water and diluted NaCl solution, then dried over $MgSO_4$, filtered and concentrated to yield α-dihydroionol (2.78 g, 94%), which was dissolved in CH_2Cl_2 (20 ml). To the solution was added $Ca(OCl)_2$ (2.7 g, 18.8 mmol, 1.3 equiv.) and the mixture was cooled to 5°C, then KH_2PO_4 (6.24 g) in water (30 ml) was added dropwise. The mixture was stirred for 30 h at RT, then poured in water (100 ml), and workup was effected as described above to yield a clear, colourless liquid (2.95 g) which was purified by FC (heptane/MTBE 30:1) to yield 4-(5-chloro-2,2-dimethyl-6-methylenecyclohexyl)butan-2-ol (0.90 g, 28%). This product (0.60 g) was dissolved in THF (20 ml) and AcOH (2.3 g) followed by zinc powder (2.60 g) were added. The mixture was stirred at RT for 21 h, then poured in 2M aq. NaOH-solution (100 ml) and further worked up as described above to yield 4-(2,2-dimethyl-6-methylenecyclohexyl)butan-2-ol (0.6 g) as a colourless oil. This product was dissolved in CH_2Cl_2 (50 ml) and pyridinium chlorochromate (PCC, 0.92 g, 4.3 mmol, 1.2 equiv.) was added. The mixture was stirred stirred for 3 h at RT. The mixture was filtered over silica gel, the filter cake was rinsed with MTBE, and the filtrate was dried. The residue was purified by FC (heptane/MTBE 95:5) to yield rac.-**5** as a colourless oil (300 mg, 46%, purity according to GC-MS 91%, the remainder being 7% of α-dihydroionone and 2% of β-dihydroionone). NMR and MS spectra were in accordance with published data.[18] Chiral GC-analysis (Astec Chiraldex G-DP): (S)-5 36.34 min, (R)-5 36.62 min.

*10.13. Preparation of (R)-**5** and (S)-**5** by preparative chiral HPLC*

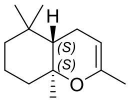

(*R*)-**5**


(*S*)-**5**

In order to provide an independent reference of both enantiomers of **5**, racemic **5** as prepared in chapter 10.12. was subjected to preparative chiral HPLC (Daicel, Chiralpak IG amylose-based, 5 µm 10*250mm, flow: 5.6 ml/min, column Temp: 15 °C, isocratic n-Hexane/MTBE 95:5, 20 min, injection of 10 µl of 100 mg/ml sample in n-hexane, (R)-(-)-**5** 11.87 min, (S)-(+)-**5** 13.71 min). From the cumulated fractions of 30 runs, the following samples were isolated after removal of the solvents.

(*R*)-(-)-**5** (16 mg, purity 85% by GC-MS) $[\alpha]_D$ = -17.7 (c = 0.68, $CHCl_3$). Chiral GC-analysis (Astec Chiraldex G-DP): 36.68 min, 98% e.e.
(*S*)-(+)-**5**, (12 mg, purity 98% by GC-MS) $[\alpha]_D$ = +16.4 (c = 0.59, $CHCl_3$). Chiral GC-analysis (Astec Chiraldex G-DP): 36.39 min, 95% e.e.

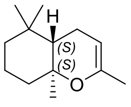
*10.14. (4aS,8aS)-2,5,5,8a-tetramethyl-4a,5,6,7,8,8a-hexahydro-4H-chromene (S,S)-**4** with AacSHC 215G2*


(*S*,*S*)-**4**

The MTBE-extract of the biocatalytic conversion of geranylacetone (*E*-**2**, 2.1 g, 10.8 mmol) with *Aac*SHC 215G2 (250 g/l, 300 ml, cf. chapter 9.3.) contained 79% of product according to GC-FID with external calibration using a racemic product standard. The extract was concentrated in a rotary evaporator, and the residue was purified via column chromatography (pentane/MTBE 60:1; the product is unstable on silica gel and partially decomposes). The combined fractions were concentrated, yielding the product (*S*,*S*)-**4** as a colourless oil (250 mg, 12%, purity by GC/MS 99.4%). $[\alpha]_D$ = -19.0 (c = 1.35, $CHCl_3$). Chiral GC analysis (Hydrodex-beta-3P): 37.70 min, >99.9% e.e.

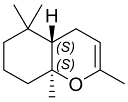
^{1}H-NMR ($CDCl_3$, 400 MHz): 4.46 (br. dd, *J*=5.3, 0.9 Hz, 1 H), 1.88 - 1.97 (m, 1 H), 1.80 - 1.86 (m, 1 H), 1.72 - 1.80 (m, 1 H), 1.68 - 1.71 (m, 3 H), 1.51 - 1.64 (m, 2 H), 1.39 - 1.51 (m, 3 H), 1.25 - 1.33 (m, 1 H), 1.18 (s, 3 H), 0.92 (s, 3 H), 0.82 (s, 3 H). ^{13}C-NMR ($CDCl_3$, 101 MHz): 148.0 (s), 95.0 (d), 76.4 (s), 48.4 (d), 41.6 (t), 40.0 (t), 33.2 (s), 32.2 (q), 20.8 (q), 20.5 (q), 19.8 (t), 19.2 (t), 19.1 (q). GC-MS (EI, 70 eV): 194 (17), 179 (6), 161 (14), 151 (6), 136 (12), 123 (27), 109 (100), 95 (21), 81 (25), 71 (19), 55 (21), 43 (61). HR-MS (ESI, positive mode): $C_{13}H_{22}O$ calcd. for $[M+H]^+$ 195.1743, found 195.1744.

*10.15. (4aS,8aS)-2,5,5,8a-tetramethyl-4a,5,6,7,8,8a-hexahydro-4H-chromene (S,S)-**4** with AciSHC_R2.1*


(*S*,*S*)-**4**

The EtOAc-extract of the biocatalytic conversion of geranylacetone (*E*-**2**, 2.4 g, 12.3 mmol) with *Aci*SHC_R2.1 (400 ml reaction, 40 mM substrate, 72 hours at 30 °C) contained 58% of product according to GC-FID. The product was purified as described above in chapter 10.14. to yield (*S*,*S*)-**4** (100 mg, 13%, purity by GC-MS 98%). NMR and MS data were identical to the product isolated in chapter 10.14 Chiral GC analysis (Hydrodex-beta-3P): 37.70 min, >99.9% e.e.

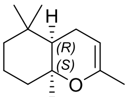
*10.16. Rac.-(4aS,8aS)-2,5,5,8a-tetramethyl-4a,5,6,7,8,8a-hexahydro-4H-chromene (rac. (S,S)-**4**)*



rac.-(*S*,*S*)-**4**

The solution of (*E*)-**2** (2.00 g, 10.3 mmol) in CH_2Cl_2 (40 ml) was cooled to -78 °C, then fluorosulfonic acid (0.65 ml, 11.3 mmol, 1.1 equiv) was added dropwise under stirring. Stirring was continued at -78 °C for 45 min, then at -50 °C for 30 min. Additional fluorosulfonic acid ((0.65 ml, 11.3 mmol, total 2.2 equiv) were added and stirring was continued at -50 °C for 30 min. The solution was poured in 2M aq. NaOH-solution (100 ml) and extracted with MTBE (100 ml). The organic layer was washed with water and brine and dried over $MgSO_4$. The product was purified as described in chapter 10.13 to yield rac.-(S,S)-**4** as a colourless oil (290 mg, 15%, purity according to GC-MS 96%). NMR and MS data were identical to the product in chapter 10.13. Chiral GC-analysis (Hydrodex-beta-3P): (*S*,*S*)-(-)-**4** (37.72 min), (*R*,*R*)-(+)-**4** (38.27 min).

*10.17. (4aR,8aS)-2,5,5,8a-tetramethyl-4a,5,6,7,8,8a-hexahydro-4H-chromene (R,S)-**4** with AacSHC 215G2*

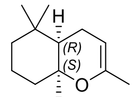


(4a*R*,8a*S*)-**4**

The MTBE-extract of the biocatalytic conversion of nerylacetone (*Z*-**2**, 525 mg, 2.2 mmol) with *Aac*SHC 215G2 (250 g/l, 350 ml, cf. chapter 9.3.) contained 60% of product according to GC-FID with external calibration using a racemic product standard. The extract was concentrated in a rotary evaporator and the residue was purified via column chromatography (pentane/MTBE 60:1; the product is unstable on silica gel and partially decomposes. After careful removal of the solvent the product (4a*R*,8a*S*)-**4** was obtained as a colourless oil (80 mg, 15%, purity by GC/MS 99.5%).
$[\alpha]_D$ = -37.8 (c = 1.02, $CHCl_3$). Chiral GC (Hydrodex-beta-3P): 30.76 min, >99.9% e.e. The absolute configuration was assigned tentatively based on the observed trend of *Aac*SHC 215G2 to yield 4a*S*-configurated products from *Z*-substrates.

1H-NMR (CDCl3, 400 MHz): 4.38 (ddt, J=5.1, 2.1, 1.0, 1.0 Hz, 1 H), 2.15 - 2.24 (m, 1 H), 1.72-1.99 (m, 3 H), 1.67 (dt, J=2.2, 1.3 Hz, 3 H), 1.15 - 1.44 (m, 5 H), 1.17 (s, 3 H), 0.89 (s, 3 H), 0.86 (s, 3 H). 13C-NMR (CDCl3, 101 MHz): 148.7 (s), 94.5 (d), 74.7 (s), 44.0 (d), 42.0 (t), 39.6 (t), 33.7 (s), 32.5 (q), 26.5 (q), 21.2 (q), 20.5 (q), 19.8 (t), 18.1 (t). GC-MS (EI, 70 eV): 194 (7), 179 (6), 151 (6), 136 (6), 124 (14), 109 (100), 95 (8), 81 (11), 71 (14), 55 (12), 43 (27). HR-MS (ESI, positive mode): C13H22O calcd. for [M+H]+ 195.1743, found 195.1743.

10.18. Rac.-(4aR,8aS)-2,5,5,8a-tetramethyl-4a,5,6,7,8,8a-hexahydro-4H-chromene (rac. (4aR,8aS)-4)



rac. (4a*R*,8a*S*)-**4**

The procedure described in chapter 10.16. was repeated with (*Z*)-**2** (4.0 g) to yield rac.-(4a*R*,8a*S*)-**4** as a colourless oil (750 mg, 19%, purity according to GC-MS 97%). NMR and MS data were identical to the product in chapter 10.17. Chiral GC-analysis (Hydrodex-beta-3P): (4a*R*,8a*S*)-(-)-**4** (29.56 min), (4a*S*,8a*R*)-(+)-**4** (30.77 min) (attribution of absolute configuration cf. chapter 10.17.).

B. Supplementary Figures

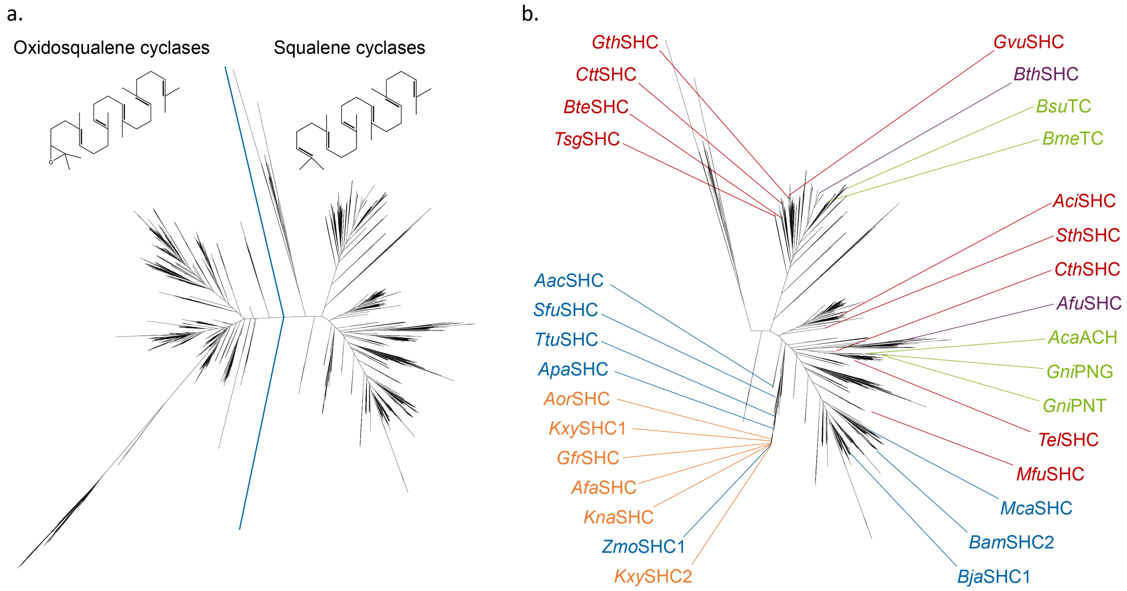


Figure S1: a) An approximate maximum likelihood tree of all 5628 sequences in Uniprot (database accessed on 05.03.2018) containing the two PFAM domains PFAM13249 PFAM13243 found in all SHCs. The tree is split into oxidosqualene cyclases (OSCs) of mainly eukaryotic origin and squalene-hopene cyclases (SHCs) of mainly prokaryotic origin. Also shown are the two standard substrates for enzymes of the respective families **b)** Close-up of the part of the phylogenetic tree containing SHCs and the location of the 31 SHCs investigated in this study. These span most major branches of the phylogenetic tree. The SHCs are colour coded. Blue: Bacterial SHCs converting squalene into hopene and hopanol. Green: Plant and *Bacillus* SHCs catalyzing unusual cyclization of squalene. Red: Novel SHC homologs from thermophilic bacteria. Orange: Novel SHC homologs from the clade, which includes the promiscuous ZmoSHC1 variant. Purple: Novel uncharacterized SHC homologs.

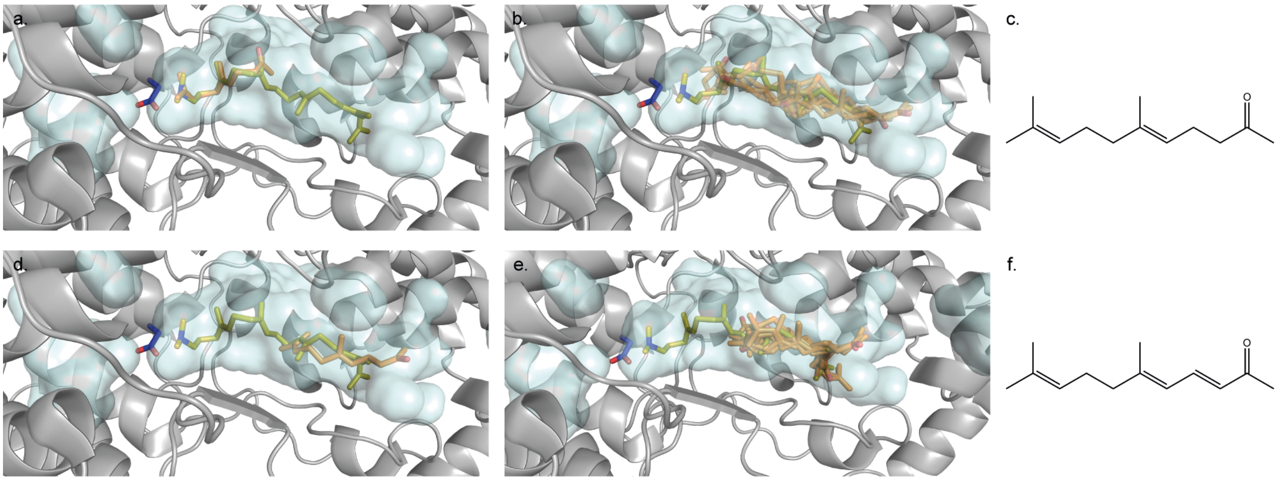
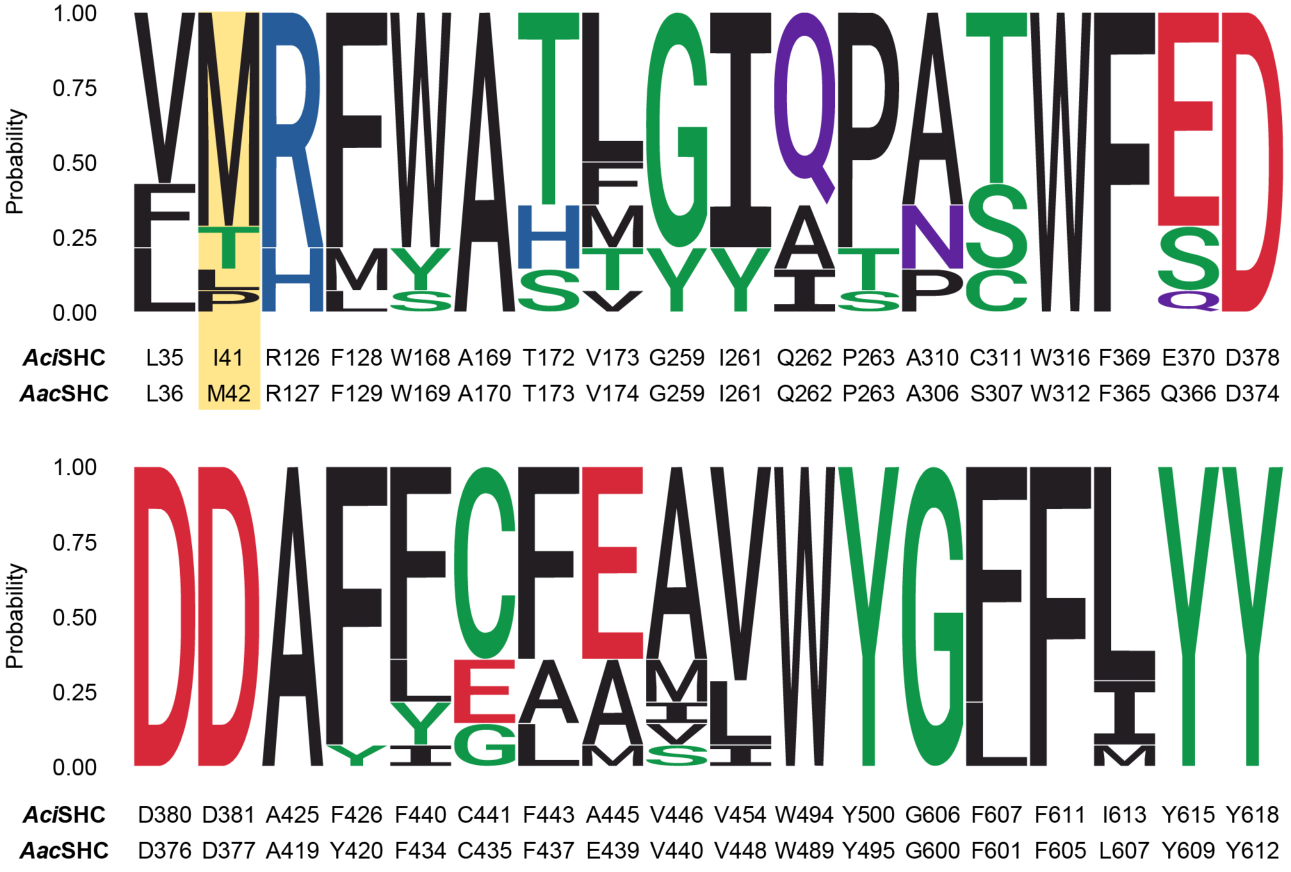


Figure S2: Docking of geranylacetone ((*E*)-**2**) **a)**: best pose, **b)**: 9 best poses and (*E*)-pseudoionone ((*E*)-**1**) **d)** best pose, **e)**: 9 best poses into the active pocket of *Aac*SHC (1UMP). The docked molecules are shown in orange, the cocrystallized 2-azasqualene in yellow and the catalytic acid D376 in dark blue. For geranylacetone ((*E*)-**2**), the best docking mode has a very similar pre folding to 2-azasqualene, and the terminal double bond is at an optimal distance of 2.7 Å from D376 for protonation. For (*E*)-pseudoionone ((*E*)-**1**) all nine docking modes are at a distance of more than 8 Å from D376. Therefore, the higher rigidity of **3** due to the additional conjugated γ,δ-double bond might render a productive pre-folding impossible. **c)** Structure of geranylacetone and **f**) (*E*)-pseudoionone.



***Aci*SHC**	L35	I41	R126	F128	W168	A169	T172	V173	G259	I261	Q262	P263	A310	C311	W316	F369	E370	D378
***Aac*SHC**	L36	M42	R127	F129	W169	A170	T173	V174	G259	I261	Q262	P263	A306	S307	W312	F365	Q366	D374

***Aci*SHC**	D380	D381	A425	F426	F440	C441	F443	A445	V446	V454	W494	Y500	G606	F607	F611	I613	Y615	Y618
***Aac*SHC**	D376	D377	A419	Y420	F434	C435	F437	E439	V440	V448	W489	Y495	G600	F601	F605	L607	Y609	Y612

Figure S3: A sequence logo of the active pocket residues [19] of the 14 SHCs converting (*E*/*Z*)-geranylacetone (**2**) into products **4** and **12**. Aligned below are the sequences of *Aci*SHC, which additionally catalyzes the formation of monocyclic products **5** and **10** and *Aac*SHC, as a reference for the usual amino acid numbering of SHCs. Highlighted in yellow is the site for which *Aci*SHC displayse an unique amino acid residue compared to all other SHCs with activity towards **2**.

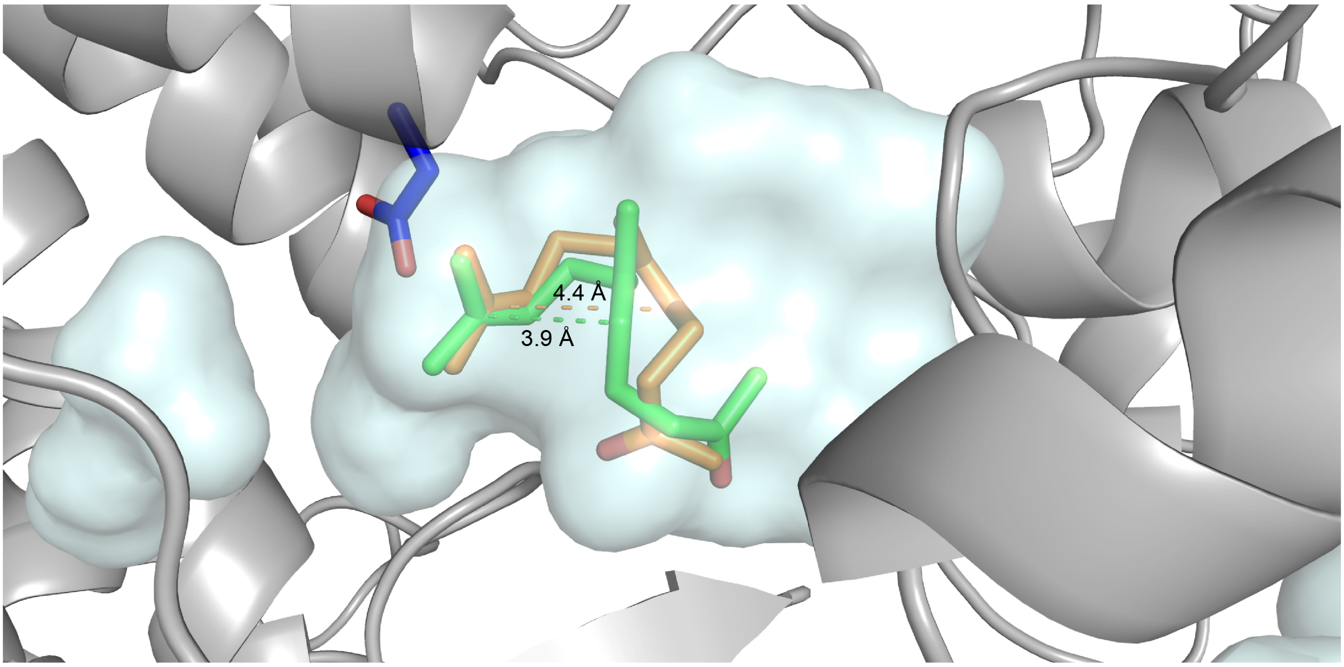


Figure S4: Docking of geranylacetone ((*E*)-**2**) (orange) and nerylacetone ((*Z*)-**2**) (green) via AutoDock Vina into a homology model of *Aci*SHC which was prepared by Swiss model. The corresponding distances for the cyclization reaction are shown. The catalytic aspartate D380 is shown in blue.

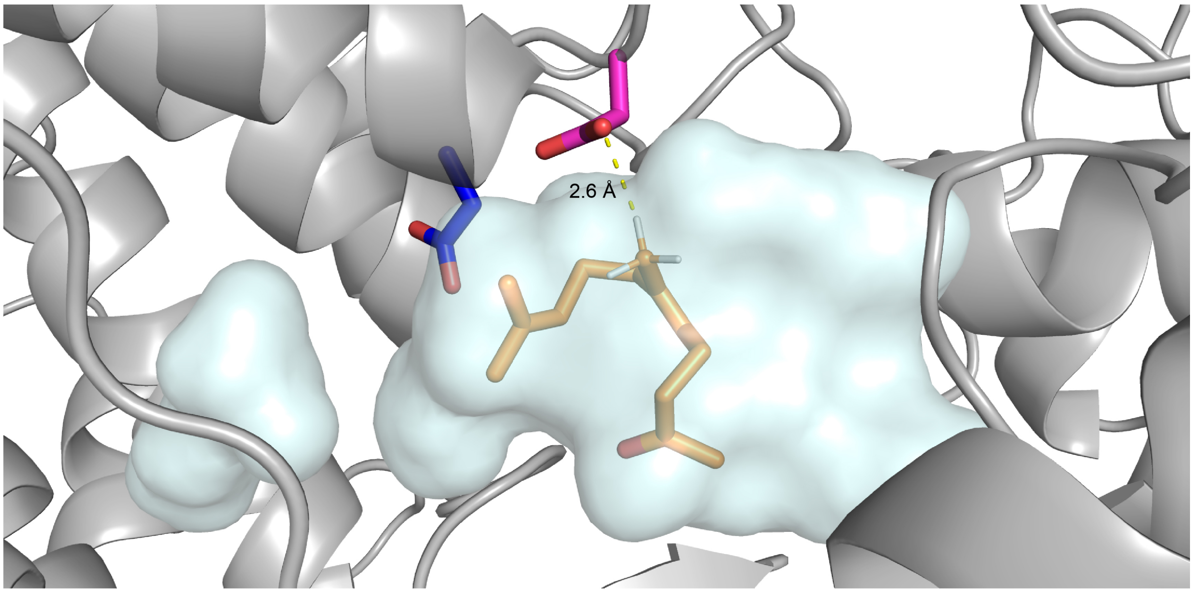


Figure S5: Docking of geranylacetone ((*E*)-**2**) (orange) into a homology model of *Aci*SHC. The distance from the hydrogen of the exocyclic methyl-group of (*E*)-**2** to D378 is shown. The distance of 2.6 Å is within the van der Waals distance between hydrogen and carbon (<2.9 Å) and therefore allows for proton transfer.[20] Therefore, D378 (purple) might be responsible for deprotonation of the monocyclic carbocation intermediate, resulting in the exomethylene containing γ-dihydroionone (**5**). The catalytic aspartate D380 is shown in blue.

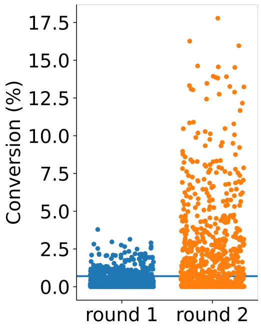


Figure S6: Conversion of (*E*/*Z*)-geranylacetone (**2**) to γ-dihydroionone (**5**) (%FID) by the SHC variants generated in the first and second round of evolution.

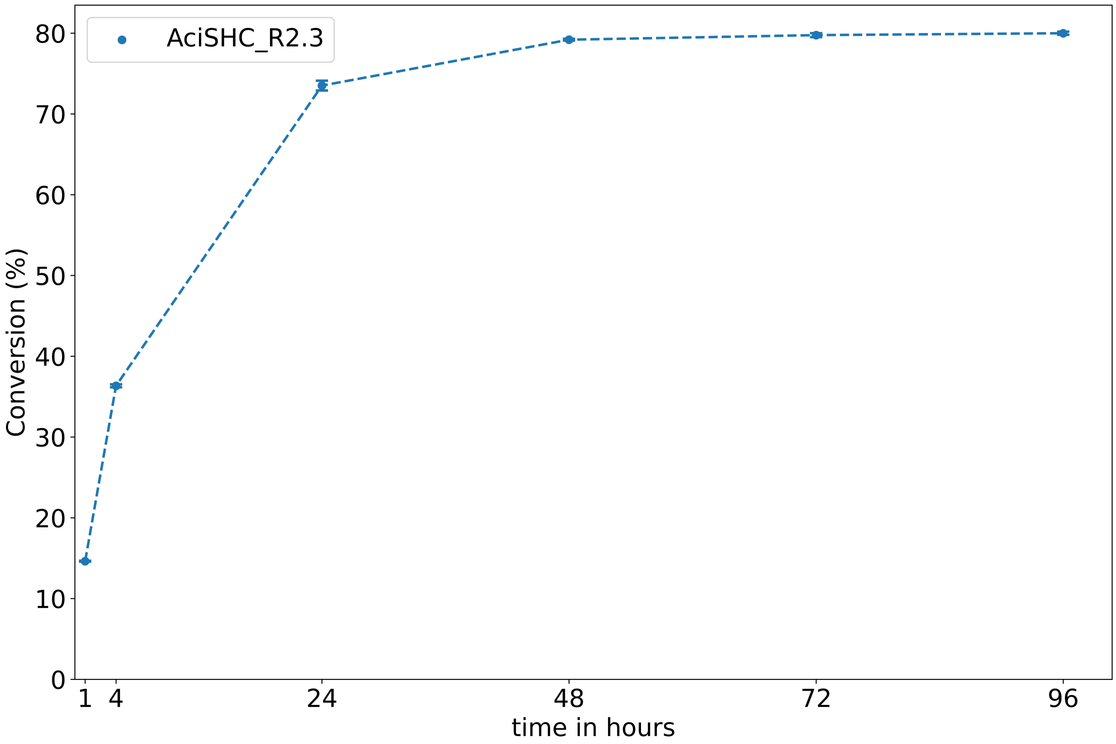


Figure S7: Time course of *Aci*SHC_R2.3. The experiments were conducted at 40 °C with an OD of 120 for 1,4,24,48,72,96 hours in triplicates. We employed 10 mM (*Z*)-**2** as substrate and measured the conversion in % FID to γ-dihydroionone ((*R*)-**5**). The average total recovery was 102±5%.

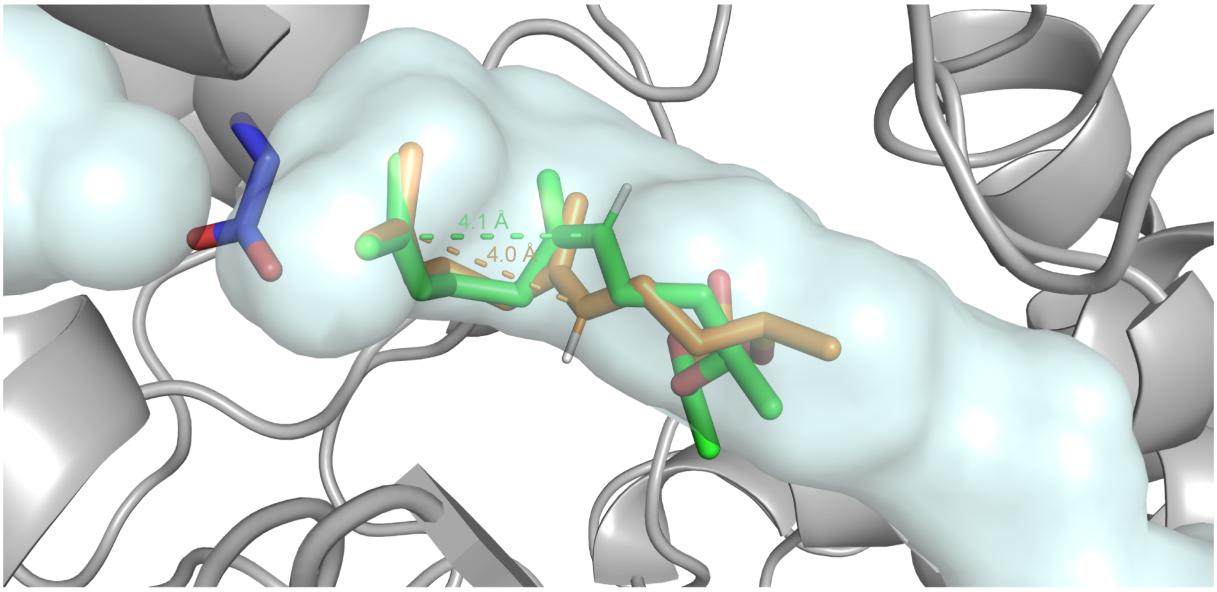


Figure S8: Docking of *(E)*-tangerinol ((*E*)-**14**) (orange) and (*Z*)-tangerinol ((*Z*)-**14**) (green) via AutoDock Vina into a homology model of *Aac*SHC 215G2 prepared by Swiss model. The corresponding distances for the cyclization reaction are shown. The catalytic aspartate D376 is shown in blue.

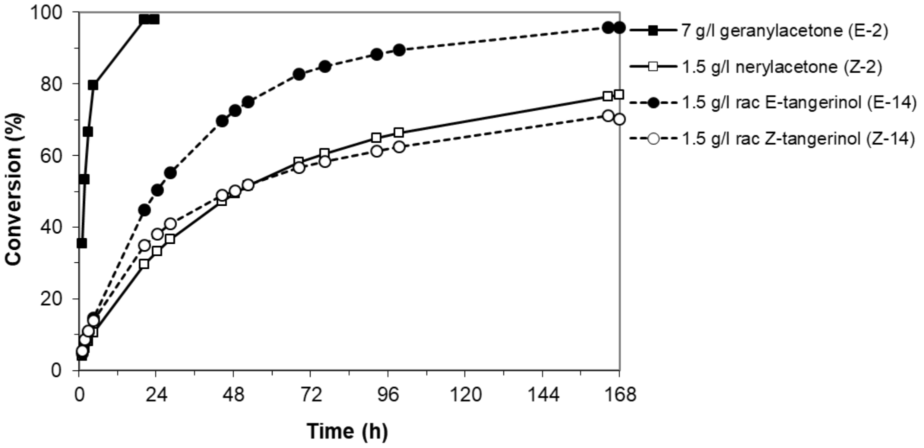
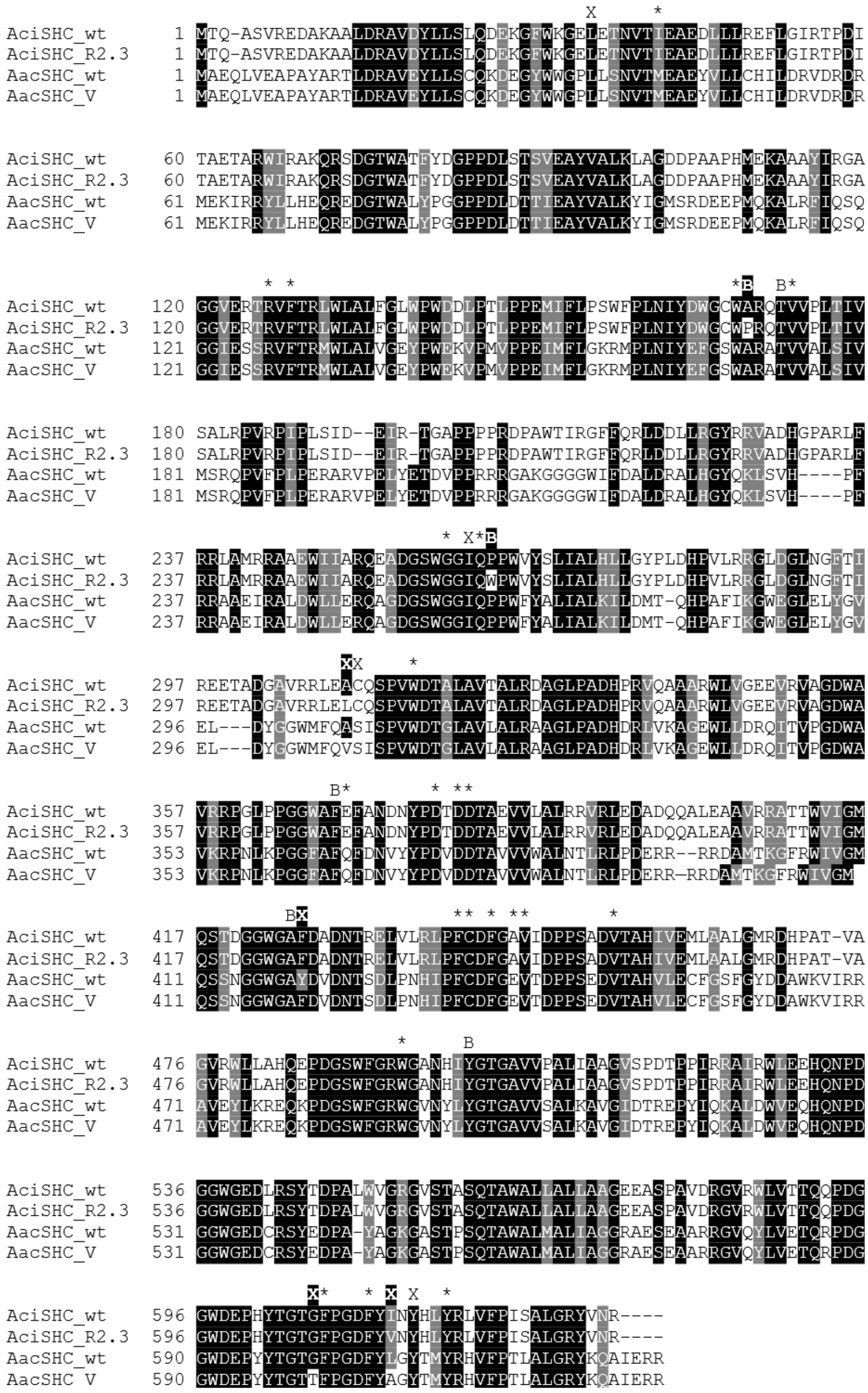


Figure S9: Preparative scale cyclization of geranylacetone (**2**) and tangerinol (**14**) with *Aac*SHC 215G2 (250 g/l cells wet weight). Geranylacetone (**2**): plain lines, racemic tangerinol (**14**): dotted lines. Almost full conversion (95.6%) was obtained after 168 h with racemic *E*-tangerinol (*E*-**14**). Conversion with racemic *Z*-tangerinol (*Z*-**14**) was only 71.0% in the same time. Geranylacetone (*E*-**2**) was fully converted in 24 h. Conversion of nerylacetone (*Z*-**2**) was 76% after 168h.



```
                                                    X     *
AciSHC_wt      1 MTQ-ASVREDAKAALDRAVDYLLSLQDEKGFWKGELETNVTIEAEDLLLREFLGIRTPDI
AciSHC_R2.3    1 MTQ-ASVREDAKAALDRAVDYLLSLQDEKGFWKGELETNVTIEAEDLLLREFLGIRTPDI
AacSHC_wt      1 MAEQLVEAPAYARTLDRAVEYLLSCQKDEGYWWGPLLSNVTMEAEYVLLCHILDRVDRDR
AacSHC_V       1 MAEQLVEAPAYARTLDRAVEYLLSCQKDEGYWWGPLLSNVTMEAEYVLLCHILDRVDRDR

AciSHC_wt     60 TAETARWIRAKQRSDGTWATFYDGPPDLSTSVEAYVALKLAGDDPAAPHMEKAAAYIRGA
AciSHC_R2.3   60 TAETARWIRAKQRSDGTWATFYDGPPDLSTSVEAYVALKLAGDDPAAPHMEKAAAYIRGA
AacSHC_wt     61 MEKIRRYLLHEQREDGTWALYPGGPPDLDTTIEAYVALKYIGMSRDEEPMQKALRFIQSQ
AacSHC_V      61 MEKIRRYLLHEQREDGTWALYPGGPPDLDTTIEAYVALKYIGMSRDEEPMQKALRFIQSQ

                       * *                                       *B  B*
AciSHC_wt    120 GGVERTRVFTRLWLALFGLWPWDDLPTLPPEMIFLPSWFPLNIYDWGCWARQTVVPLTIV
AciSHC_R2.3  120 GGVERTRVFTRLWLALFGLWPWDDLPTLPPEMIFLPSWFPLNIYDWGCWPRQTVVPLTIV
AacSHC_wt    121 GGIESSRVFTRMWLALVGEYPWEKVPMVPPEIMFLGKRMPLNIYEFGSWARATVVALSIV
AacSHC_V     121 GGIESSRVFTRMWLALVGEYPWEKVPMVPPEIMFLGKRMPLNIYEFGSWARATVVALSIV

AciSHC_wt    180 SALRPVRPIPLSID--EIR-TGAPPPPRDPAWTIRGFFQRLDDLLRGYRRVADHGPARLF
AciSHC_R2.3  180 SALRPVRPIPLSID--EIR-TGAPPPPRDPAWTIRGFFQRLDDLLRGYRRVADHGPARLF
AacSHC_wt    181 MSRQPVFPLPERARVPELYETDVPPRRRGAKGGGGWIFDALDRALHGYQKLSVH----PF
AacSHC_V     181 MSRQPVFPLPERARVPELYETDVPPRRRGAKGGGGWIFDALDRALHGYQKLSVH----PF

                                       * X*B
AciSHC_wt    237 RRLAMRRAAEWIIARQEADGSWGGIQPPWVYSLIALHLLGYPLDHPVLRRGLDGLNGFTI
AciSHC_R2.3  237 RRLAMRRAAEWIIARQEADGSWGGIQWPWVYSLIALHLLGYPLDHPVLRRGLDGLNGFTI
AacSHC_wt    237 RRAAEIRALDWLLERQAGDGSWGGIQPPWFYALIALKILDMT-QHPAFIKGWEGLELYGV
AacSHC_V     237 RRAAEIRALDWLLERQAGDGSWGGIQPPWFYALIALKILDMT-QHPAFIKGWEGLELYGV

                              XX    *
AciSHC_wt    297 REETADGAVRRLEACQSPVWDTALAVTALRDAGLPADHPRVQAAARWLVGEEVRVAGDWA
AciSHC_R2.3  297 REETADGAVRRLELCQSPVWDTALAVTALRDAGLPADHPRVQAAARWLVGEEVRVAGDWA
AacSHC_wt    296 EL---DYGGWMFQASISPVWDTGLAVLALRAAGLPADHDRLVKAGEWLLDRQITVPGDWA
AacSHC_V     296 EL---DYGGWMFQVSISPVWDTGLAVLALRAAGLPADHDRLVKAGEWLLDRQITVPGDWA

                             B*       * **
AciSHC_wt    357 VRRPGLPPGGWAFEFANDNYPDTDDTAEVVLALRRVRLEDADQQALEAAVRRATTWVIGM
AciSHC_R2.3  357 VRRPGLPPGGWAFEFANDNYPDTDDTAEVVLALRRVRLEDADQQALEAAVRRATTWVIGM
AacSHC_wt    353 VKRPNLKPGGFAFQFDNVYYPDVDDTAVVVWALNTLRLPDERR--RRDAMTKGFRWIVGM
AacSHC_V     353 VKRPNLKPGGFAFQFDNVYYPDVDDTAVVVWALNTLRLPDERR—RRDAMTKGFRWIVGM

                         BX             ** * **       *
AciSHC_wt    417 QSTDGGWGAFDADNTRELVLRLPFCDFGAVIDPPSADVTAHIVEMLAALGMRDHPAT-VA
AciSHC_R2.3  417 QSTDGGWGAFDADNTRELVLRLPFCDFGAVIDPPSADVTAHIVEMLAALGMRDHPAT-VA
AacSHC_wt    411 QSSNGGWGAYDVDNTSDLPNHIPFCDFGEVTDPPSEDVTAHVLECFGSFGYDDAWKVIRR
AacSHC_V     411 QSSNGGWGAFDVDNTSDLPNHIPFCDFGEVTDPPSEDVTAHVLECFGSFGYDDAWKVIRR

                                   *     B
AciSHC_wt    476 GVRWLLAHQEPDGSWFGRWGANHIYGTGAVVPALIAAGVSPDTPPIRRAIRWLEEHQNPD
AciSHC_R2.3  476 GVRWLLAHQEPDGSWFGRWGANHIYGTGAVVPALIAAGVSPDTPPIRRAIRWLEEHQNPD
AacSHC_wt    471 AVEYLKREQKPDGSWFGRWGVNYLYGTGAVVSALKAVGIDTREPYIQKALDWVEQHQNPD
AacSHC_V     471 AVEYLKREQKPDGSWFGRWGVNYLYGTGAVVSALKAVGIDTREPYIQKALDWVEQHQNPD

AciSHC_wt    536 GGWGEDLRSYTDPALWVGRGVSTASQTAWALLALLAAGEEASPAVDRGVRWLVTTQQPDG
AciSHC_R2.3  536 GGWGEDLRSYTDPALWVGRGVSTASQTAWALLALLAAGEEASPAVDRGVRWLVTTQQPDG
AacSHC_wt    531 GGWGEDCRSYEDPA-YAGKGASTPSQTAWALMALIAGGRAESEAARRGVQYLVETQRPDG
AacSHC_V     531 GGWGEDCRSYEDPA-YAGKGASTPSQTAWALMALIAGGRAESEAARRGVQYLVETQRPDG

                           X*   * X X  *
AciSHC_wt    596 GWDEPHYTGTGFPGDFYINYHLYRLVFPISALGRYVNR----
AciSHC_R2.3  596 GWDEPHYTGTGFPGDFYVNYHLYRLVFPISALGRYVNR----
AacSHC_wt    590 GWDEPYYTGTGFPGDFYLGYTMYRHVFPTLALGRYKQAIERR
AacSHC V     590 GWDEPYYTGTTFPGDFYAGYTMYRHVFPTLALGRYKQAIERR
```

Figure S10: Alignment of *Aci*SHC, *Aci*SHC_R2.3, the best variant for conversion of nerylacetone ((*Z*)-**2**) into γ-dihydroionone ((*R*)-**5**) described in this work, *Aac*SHC and *Aac*SHC_V, the best variant for the same reaction from the recent work by the group of Bernhard Hauer.[21] Active pocket residues[19] are annotated: X: residues that were targeted in both protein engineering efforts, B: residues that were only targeted in this work, *: residues that were not targeted by either work. black background: positions mutated in *Aci*SHC_2.3 and/or *Aac*SHC_V.

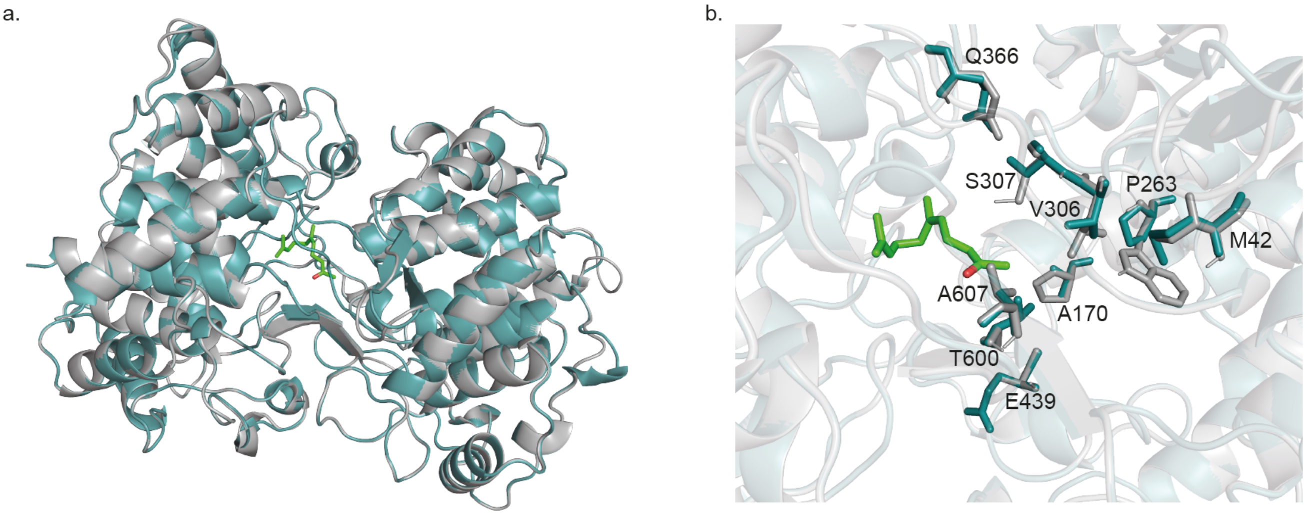


Figure S11: Alignment of homology models of *Aci*SHC_R2.3 (grey), the best variant for conversion of nerylacetone ((*Z*)-**2**) into γ-dihydroionone ((*R*)-**5**) described in this work with *Aac*SHC_V (cyan), the best variant for the same reaction from the recent work by the group of Bernhard Hauer[21]. **a)** Full homology model. **b)** Active site. Position with different amino acid residues between the two engineered enzymes are shown as sticks. Residues are labelled for *Aac*SHC_V. The docking results of (*Z*)-**2** into the homology model of *Aci*SHC_R2.3 using AutoDock Vina is shown in green. As expected, we found that the two homology models are very similar (all atom RMSD: 1.57 Å) because they are both based on the crystal structure of *Aac*SHC (PDB: 1SQC).

C. Supplementary Tables

Table S1: List of all primers used in this study.

Primer Name	Sequence
T7_fw	TAATACGACTCACTATAGGG
T7_rv	GCTAGTTATTGCTCAGCGG
L35X_rv	TTCACCTTTCCAAAAACCTTTTTC
L35X_fw	GAAAAAGGTTTTTGGAAAGGTGAANNKGAGACCAACGTTACCATTGAAG
A169X_rv	CCAACAACCCCAATCATAAATG
A169X_fw	TTTATGATTGGGGTTGTTGGNNKCGCCAGACCGTTGTGC
T172X_rv	CTGACGTGCCCAACAAC
T172X_fw	GTTGTTGGGCACGTCAGNNKGTCGTGCCGCTGACCATTG
I261X_rv	ACCACCCCAGCTACCATC
I261X_fw	GATGGTAGCTGGGGTGGTNNKCAACCTCCGTGGGTTTATAG
P263X_rv	CTGAATACCACCCCAGCTAC
P263X_fw	GTAGCTGGGGTGGTATTCAGNNKCCATGGGTTTATAGCCTGATTG
A310X_rv	TTCCAGGCGACGAACTG
A310X_fw	CAGTTCGTCGCCTGGAANNKTGCCAGAGTCCGGTTTGGG
C311X_rv	TGCTTCCAGGCGACG
C311X_fw	CGTCGCCTGGAAGCANNKCAAAGTCCGGTTTGGGATAC
F369X_rv	TGCCCAACCACCAGG
F369X_fw	CCTGGTGGTTGGGCANNKGAGTTTGCCAATGATAATTACCCG
A425X_rv	ACCCCAACCGCCATC
A425X_fw	GATGGCGGTTGGGGTNNKTTCGATGCAGATAATACCCG
F426X_rv	TGCACCCCAACCGCC
F426X_fw	GGCGGTTGGGGTGCANNKGACGCAGATAATACCCGTGAAC
Y500X_rv	AATATGATTTGCTCCCCAACG
Y500X_fw	CGTTGGGGAGCAAATCATATTNNKGGCACGGGTGCAGTTGTTC
G606X_rv	TGTACCGGTGTAATGCG
G606X_fw	CGCATTACACCGGTACANNKTTCCCGGGTGATTTCTATATTAACTATC
I613X_rv	ATAGAAATCACCCGGAAAACC
I613X_fw	GGTTTTCCGGGTGATTTCTATNNKAATTATCATCTGTATCGCCTGG
Y615X_rv	GTTAATATAGAAATCACCCGGAAAAC
Y615X_fw	GTTTTCCGGGTGATTTCTATATTAACNNKCACCTGTATCGCCTGGTGTTTC
A169_rv	CCAACAACCCCAATCATAAATG
A169WT_fw	CATTTATGATTGGGGTTGTTGGGCGCGCCAGACCGTTGTGCCGCTG
A169G_fw	CATTTATGATTGGGGTTGTTGGGGCCGCCAGACCGTTGTGCCGCTG
A169P_fw	CATTTATGATTGGGGTTGTTGGCCGCGCCAGACCGTTGTGCCGCTG
P263_rv	CTGAATACCACCCCAGCTACC
P263WT_fw	GGTAGCTGGGGTGGTATTCAGCCTCCATGGGTTTATAGCCTGATTG
P263W_fw	GGTAGCTGGGGTGGTATTCAGTGGCCATGGGTTTATAGCCTGATTG
A310_rv	TTCCAGGCGACGAACTGC
A310WT_fw	CAGTTCGTCGCCTGGAAGCATGCCAGAGTCCGGTTTGGGATAC
A310F_fw	CAGTTCGTCGCCTGGAATTTTGCCAGAGTCCGGTTTGGGATAC
A310L_fw	CAGTTCGTCGCCTGGAACTGTGCCAGAGTCCGGTTTGGGATAC
A310M_fw	CAGTTCGTCGCCTGGAAATGTGCCAGAGTCCGGTTTGGGATAC
G606_I613_rv	TGTACCGGTGTAATGCGG
G606ACC_I613VTT_fw	CCGCATTACACCGGTACAACCTTCCCGGGTGATTTCTATVTTAATTATCATCTGTATCGCCTGG
G606ACC_I613GCA_fw	CCGCATTACACCGGTACAACCTTCCCGGGTGATTTCTATGCAAATTATCATCTGTATCGCCTGG
G606KGC_I613VTT_fw	CCGCATTACACCGGTACAKGCTTCCCGGGTGATTTCTATVTTAATTATCATCTGTATCGCCTGG
G606KGC_I613GCA_fw	CCGCATTACACCGGTACAKGCTTCCCGGGTGATTTCTATGCAAATTATCATCTGTATCGCCTGG

Table S2: List of all SHCs used in this study and the plasmids, from which they are expressed. The enzymes are ordered based on their phylogenetic relationship. References on initial characterization of the enzymes are given and previously not characterized SHCs are marked with an asterisk.

SHC	Plasmid	Backbone	Source organism	NCBI Acc. Nr.	Reference
-	pET28b(+)	pET28b(+)	-	-	-
*Kna*SHC	pSHC30	pET28b(+)	*Komagataeibacter nataicola*	AQU88860.1	*
*Kxy*SHC*2*	pSHC32	pET28b(+)	*Komagataeibacter xylinus* E25	AHI26287.1	*
*Afa*SHC	pSHC27	pET28b(+)	*Acetobacter fabarum*	PAK78064.1	*
*Aor*SHC	pSHC28	pET28b(+)	*Acetobacter orleanensis* JCM 7639	GAN69910.1	*
*Kxy*SHC*1*	pSHC31	pET28b(+)	*Gluconacetobacter xylinus*	CUW48332.1	*
*Apa*SHC	pSHC13	pET28b(+)	*Acetobacter pasteurianus*	WP_012812952.1	[6]
*Gfr*SHC	pSHC29	pET28b(+)	*Gluconobacter frateurii* NBRC 103465	GAD08844.1	*
*Zmo*SHC*1*	pSHC5	pET28b(+)	*Zymomonas mobilis*	WP_011241313.1	[5]
*Sfu*SHC	pSHC16	pET28b(+)	*Syntrophobacter fumaroxidans*	WP_011698842.1	[6]
*Ttu*SHC	pSHC17	pET28b(+)	*Teredinibacter turnerae*	WP_015819476.1	[6]
*Aac*SHC	pSHC3	pET28b(+)	*Alicyclobacillus acidocaldarius*	WP_012811690.1	[4]
*Aci*SHC	pSHC8	pET28b(+)	*Acidothermus cellulolyticus*	WP_011720532.1	*
*Sth*SHC	pSHC25	pET28b(+)	*Sphaerobacter thermophilus*	WP_012872483.1	*
*Afu*SHC	pSHC12	pET28b(+)	*Aspergillus fumigatus* A1163	EDP50814.1	*
*Cth*SHC	pSHC22	pET28b(+)	*Chloracidobacterium thermophilum*	WP_014100779.1	*
*Tel*SHC	pSHC11	pET28b(+)	*Thermosynechococcus elongatus*	WP_011058142.1	*
AcaACH	pSHC18	pET28b(+)	*Adiantum capillus-veneris*	BAF93209.1	[9]
GniPNT	pSHC7	pET28b(+)	*Goniophlebium niponicum*	BAI48070.1	[10]
GniPNG	pSHC20	pET28b(+)	*Goniophlebium niponicum*	BAI48071.1	[10]
*Mfu*SHC	pSHC10	pET28b(+)	*Methylacidiphilum fumariolicum*	WP_009061034.1	*
*Bja*SHC*1*	pSHC4	pET28b(+)	*Bradyrhizobium japonicus*	CAA60250.1	[7]
*Bam*SHC*2*	pSHC14	pET28b(+)	*Burkholderia ambifaria*	ABI91648.1	[6]
*Mca*SHC	pSHC15	pET28b(+)	*Methylococcus capsulatus*	WP_010960137.1	[8]
BmeTC	pSHC6	pET28b(+)	*Bacillus megaterium*	WP_013083001.1	[12]
BsuTC	pSHC19	pET28b(+)	*Bacillus subtilis*	WP_004399534.1	[11]
*Bth*SHC	pSHC1	pET28b(+)	*Bacillus thuringiensis* B 4219	AJI35613.1	*
*Tsg*SHC	pSHC26	pET28b(+)	*Thermoactinomyces sp.* Gus2-1	KFZ40906.1	*
*Bte*SHC	pSHC21	pET28b(+)	*Brevibacillus thermoruber*	WP_029099368.1	*
*Ctt*SHC	pSHC23	pET28b(+)	*Cohnella thermotolerans*	WP_027091823.1	*
*Gvu*SHC	pSHC9	pET28b(+)	*Geobacillus vulcani*	WP_031409036	*
*Gth*SHC	pSHC24	pET28b(+)	*Geobacillus thermodenitrificans*	WP_029761705.1	*

Table S3: Sequence of the ten best variants of the second round of directed evolution of *Aci*SHC for the conversion of (*E*/*Z*)-geranylacetone (**2**) towards γ-dihydroionone (**5**). Wild type residues are shown in italics, while novel residues are shown in bold. Biocatalytic reactions in screening conditions in deep-well plates were performed with 10mM (*Z*)-**2** towards **5** in triplicates.

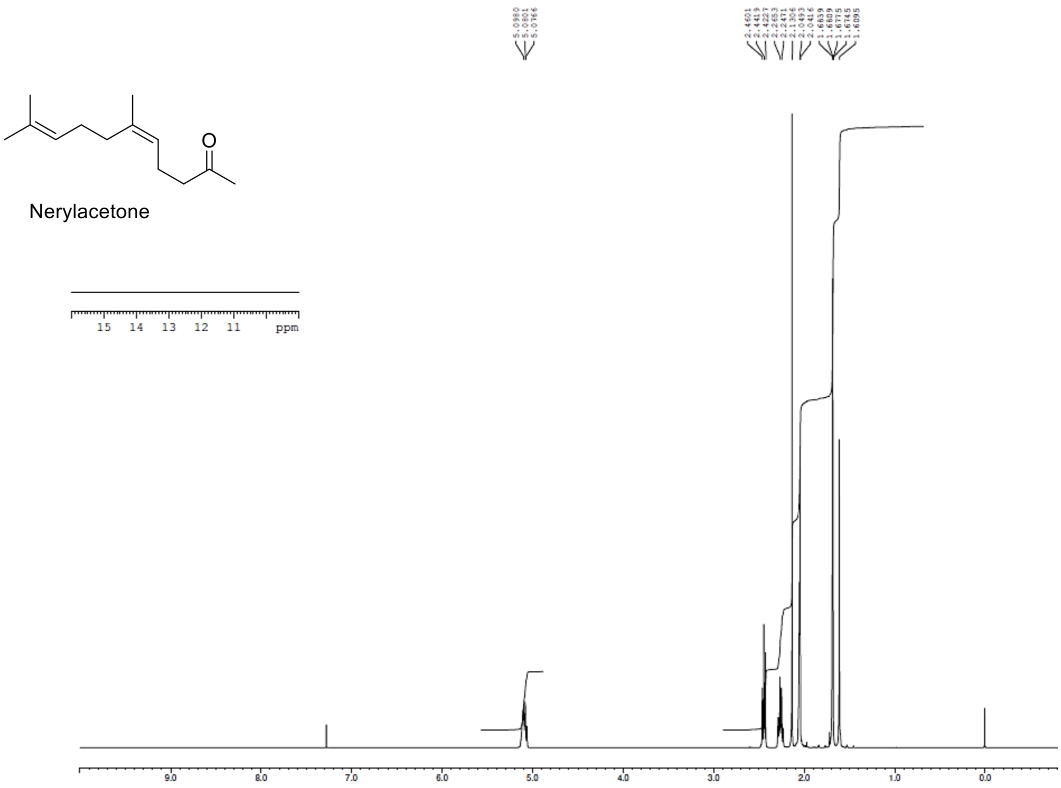
Variant	A169	P263	A310	G606	I613	Conversion (%)
*Aci*SHC_wt	*A*	*P*	*A*	*G*	*I*	0.71 ± 0.02
*Aci*SHC_R1.1	*A*	*P*	**F**	*G*	*I*	4.7 ± 0.1
*Aci*SHC_R2.1	**P**	*P*	**M**	**C**	**V**	21.4 ± 1.0
*Aci*SHC_R2.2	**P**	**W**	**M**	*G*	**L**	19.9 ± 0.3
*Aci*SHC_R2.3	**P**	**W**	**L**	*G*	**V**	18.7 ± 0.3
*Aci*SHC_R2.4	**P**	**W**	**M**	*G*	**I**	18.53 ± 0.14
*Aci*SHC_R2.5	**P**	**W**	*A*	**C**	**V**	18.0 ± 0.3
*Aci*SHC_R2.6	**P**	*P*	**L**	**C**	**V**	17.55 ± 0.02
*Aci*SHC_R2.7	**P**	**W**	*A*	**C**	**V**	17.5 ± 0.3
*Aci*SHC_R2.8	**G**	*P*	**L**	**C**	**V**	17.13 ± 0.08
*Aci*SHC_R2.9	**P**	**W**	*A*	*G*	**V**	16.9 ± 1.9
*Aci*SHC_R2.10	**G**	*P*	**F**	**C**	**V**	14.4 ± 1.7

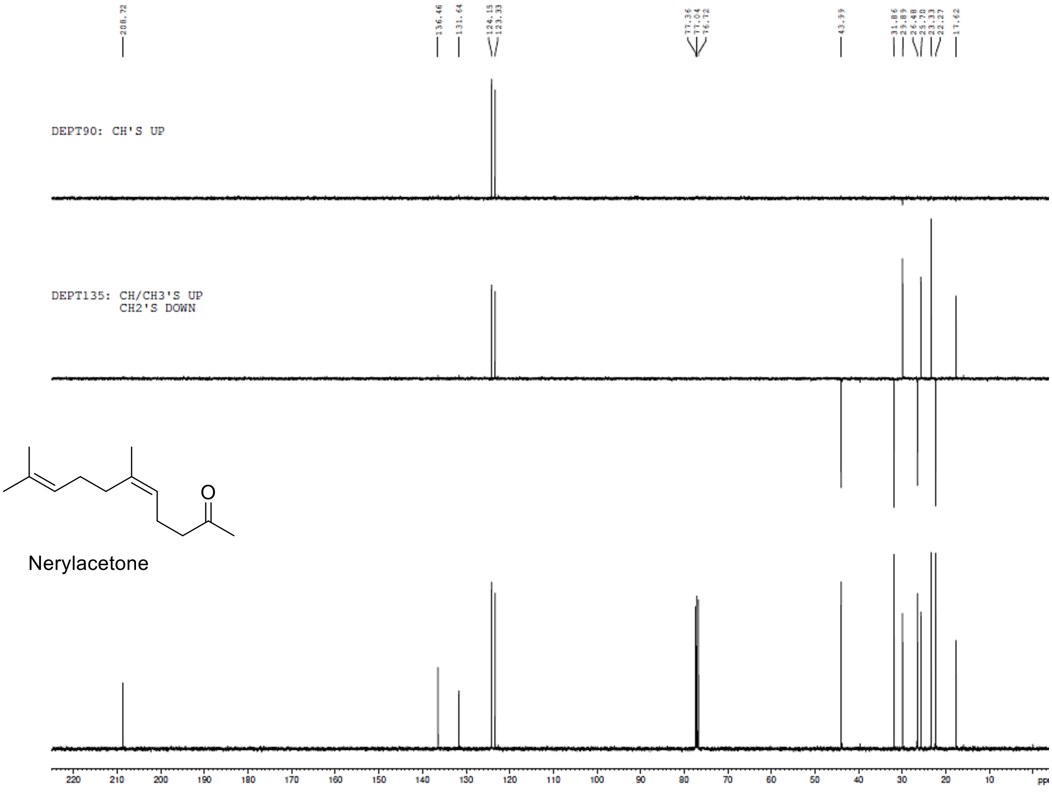
Table S4: Conversion and optical purity of biocatalytic products from (*E*/*Z*)-geranylacetone (**2**) with different variants of *Aci*SHC. Biocatalytic reactions in glass vials were performed with 10mM (*E*/*Z*)-**2** in triplicates.

Variant\Product	(*R*)-γ-dihydroionone ((R)-**5**)		(*S*,*S*)-bicyclic enol ether ((S,S)-**4**)		(*R*)-α-dihydroionone ((R)-**10**)	
	Conversion (%)	ee (%)	Conversion (%)	ee (%)	Conversion (%)	ee (%)
*Aci*SHC_wt	0.70 ± 0.04	>90%	2.34 ± 0.08	>90%	0.08 ± 0.01	n.d.
*Aci*SHC_R1.1	5.4 ± 1.1	>99%	13.0 ± 1.8	>95%	0.4 ± 0.1	>90%
*Aci*SHC_R2.1	14.5 ± 0.3	>99%	2.93 ± 0.03	>95%	1.32 ± 0.01	>90%
*Aci*SHC_R2.2	13.6 ± 1.0	99.7%	17.0 ± 1.4	>95%	0.87 ± 0.08	>90%
*Aci*SHC_R2.3	12.15 ± 0.15	> 99%	9.43 ± 0.23	>95%	1.06 ± 0.02	>90%

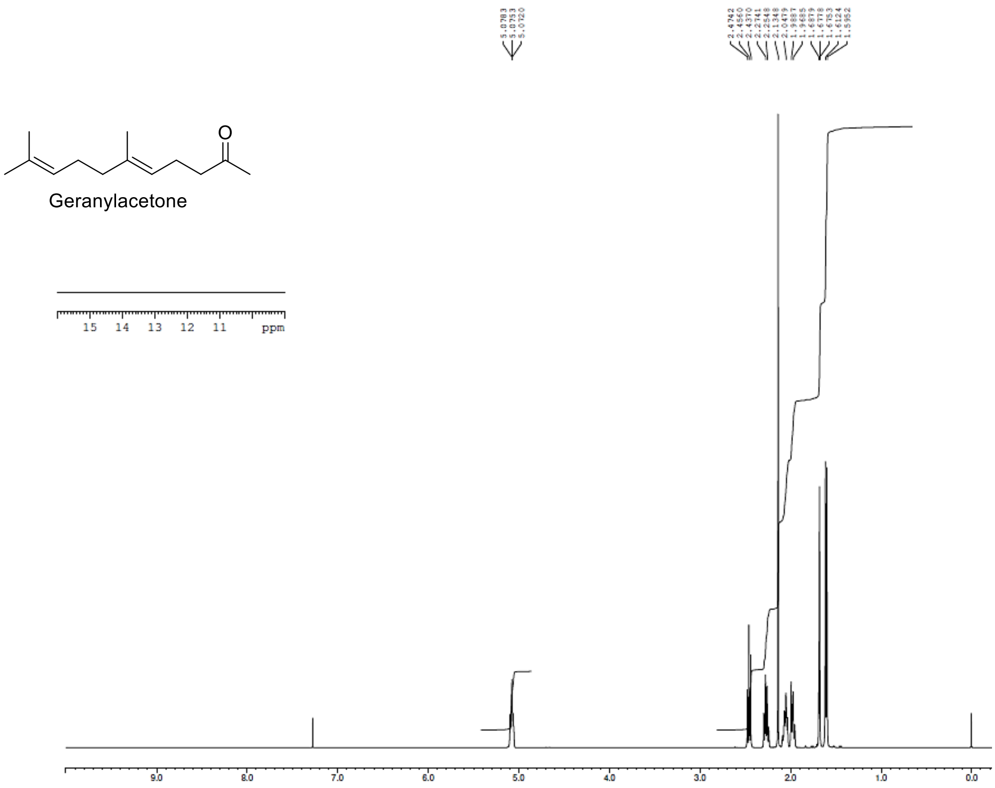
D. NMR Data

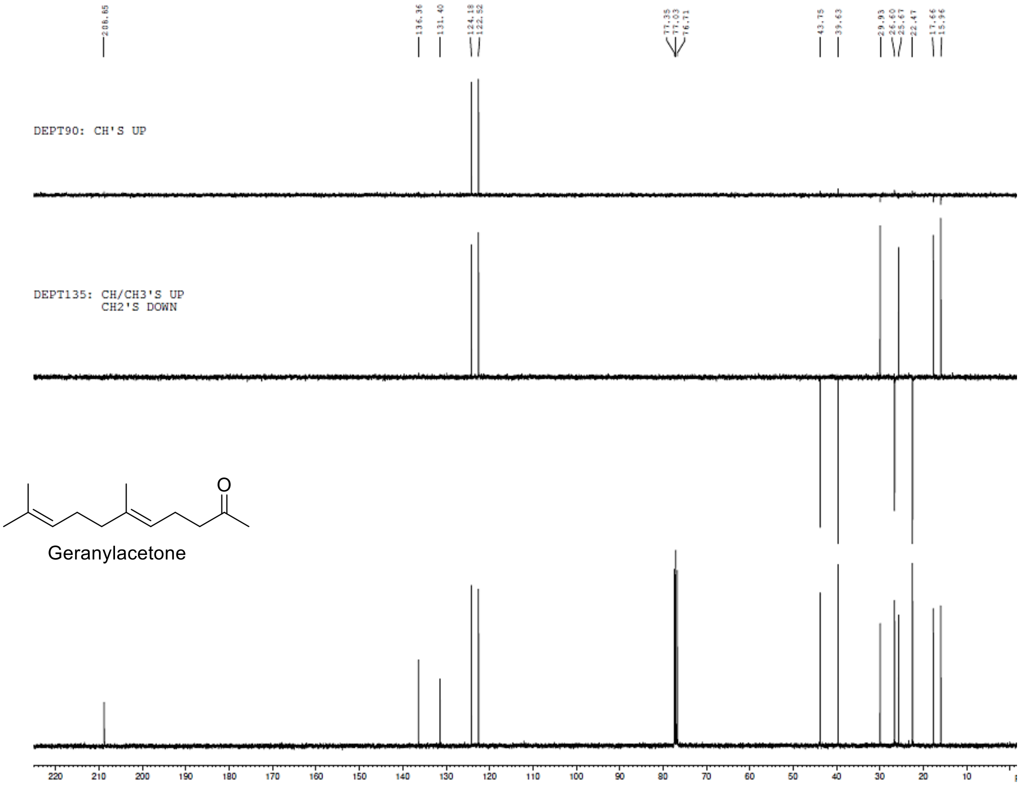
1. Nerylacetone (*Z*-2)



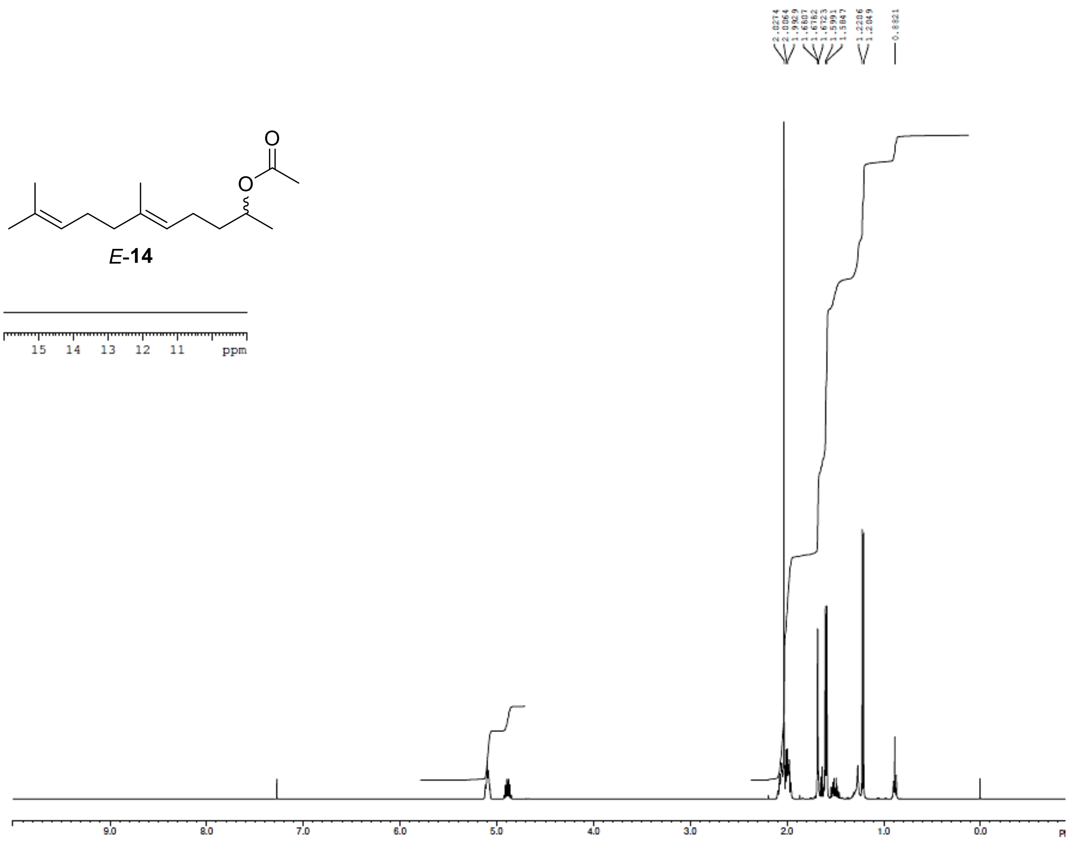


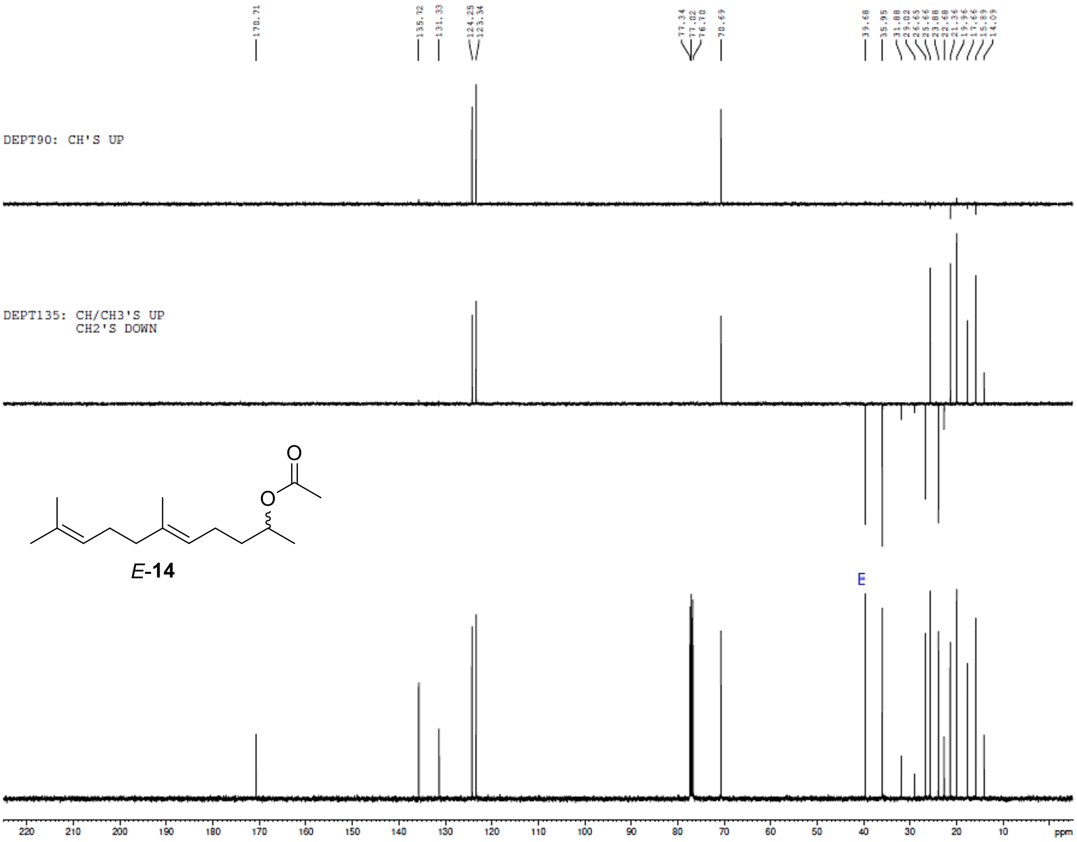
2. Geranylacetone (*E*-2)



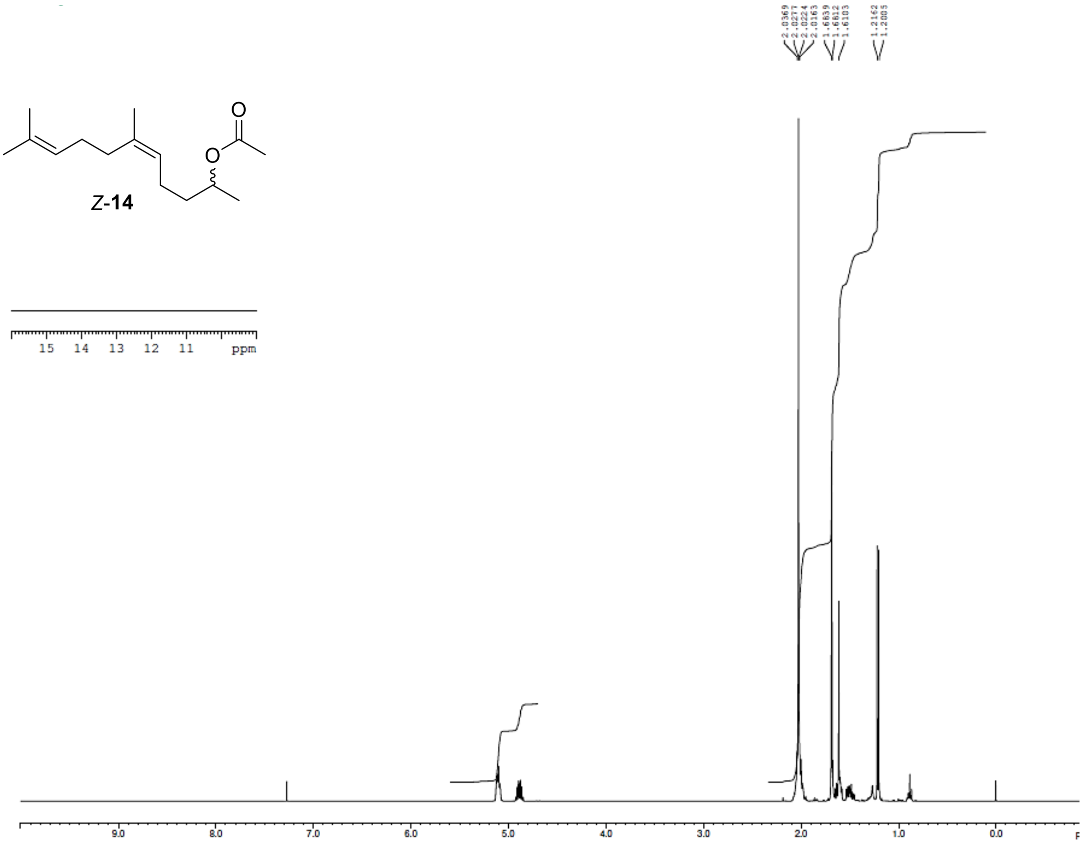


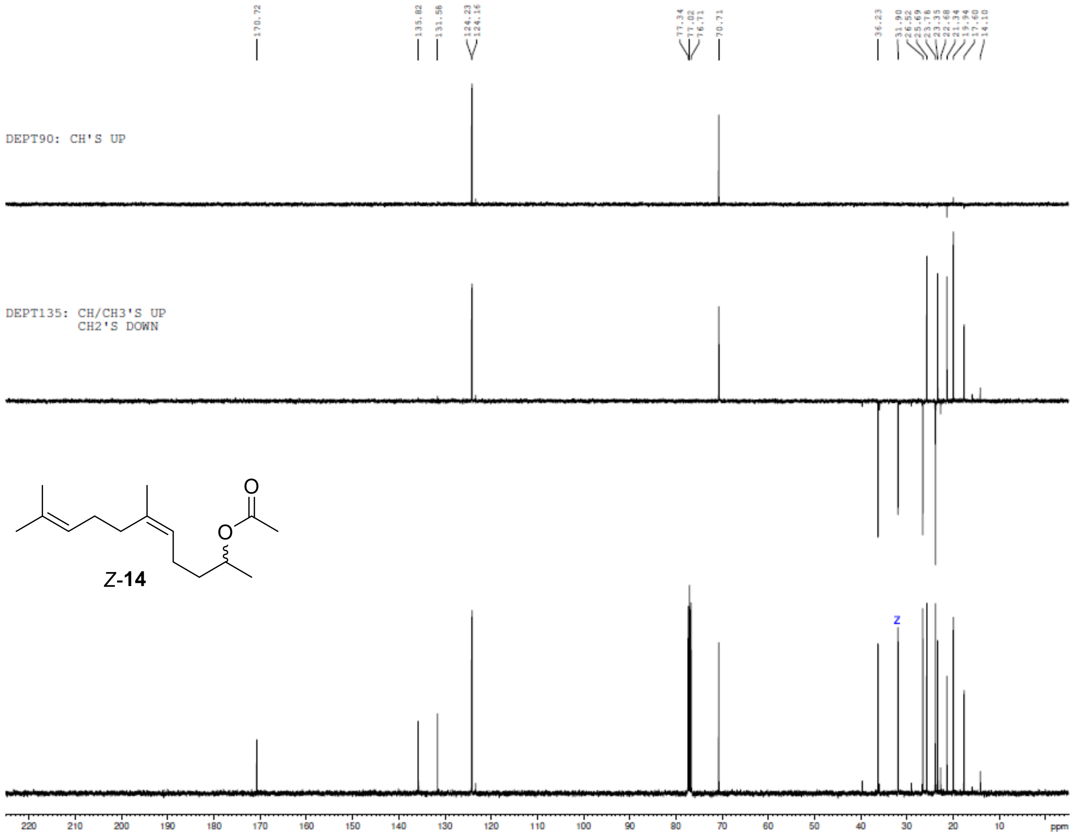
3. *E*-Tangerinol (*E*-14)



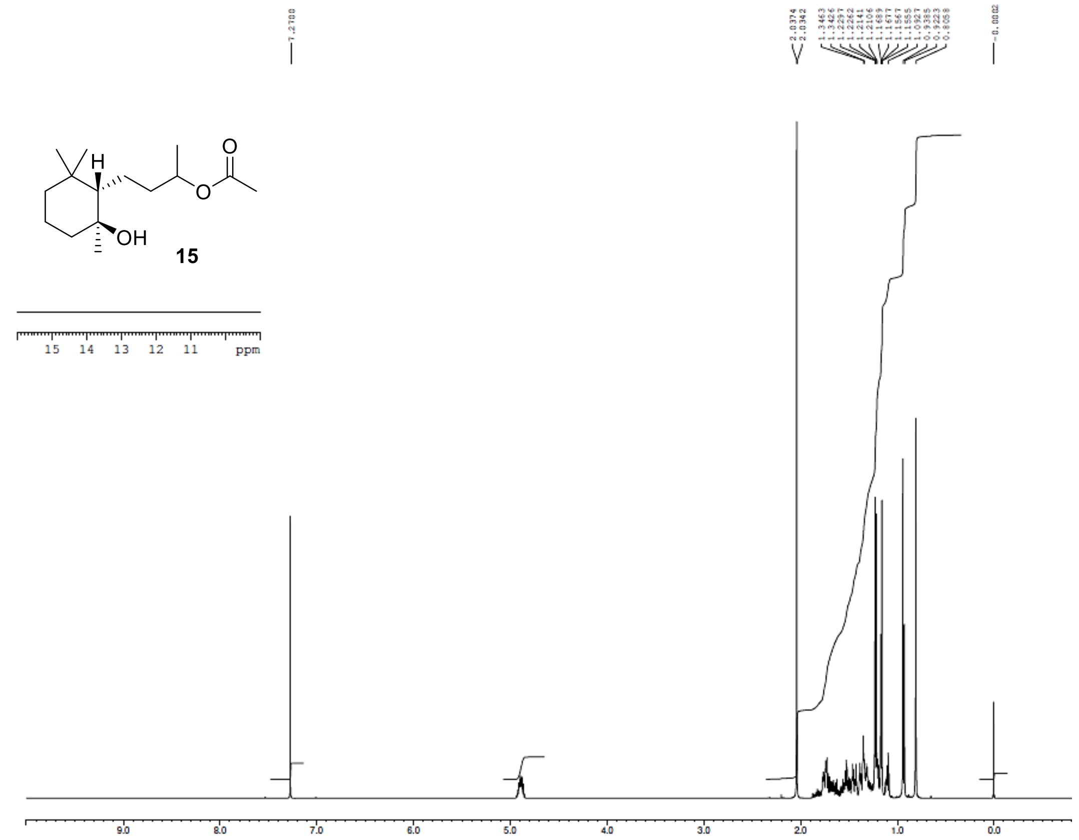


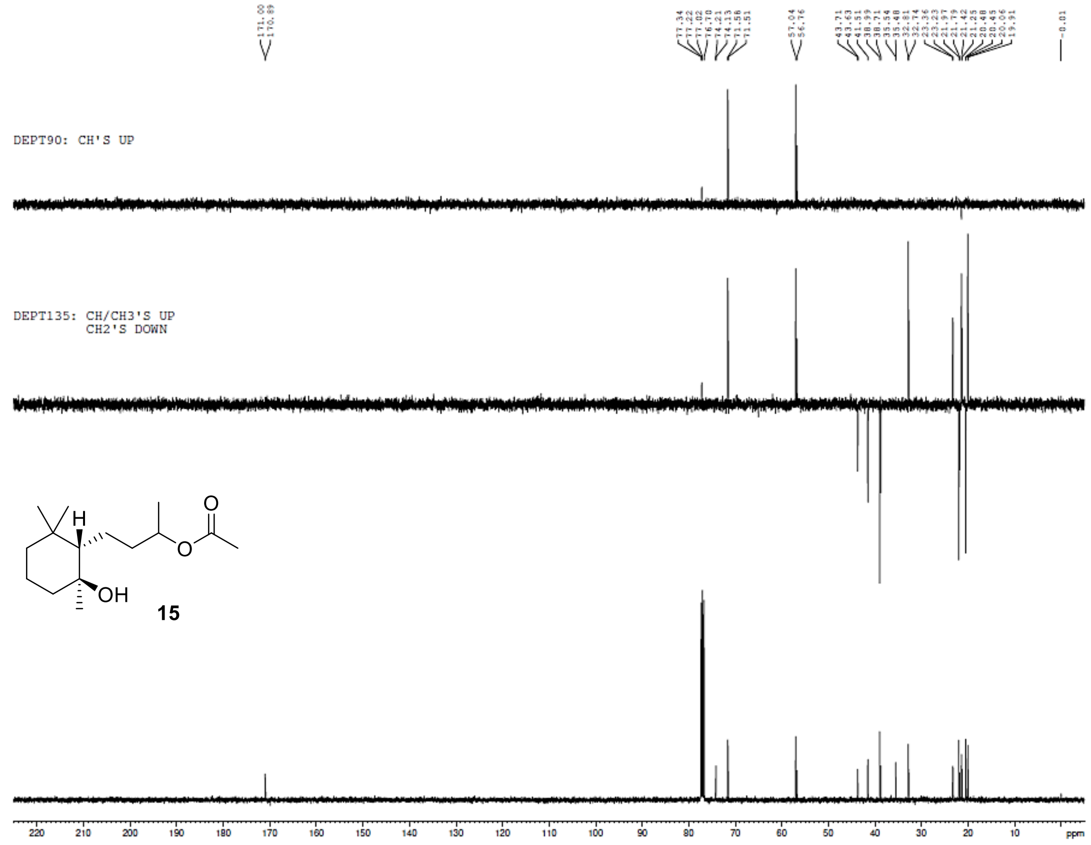
4. *Z*-Tangerinol (*Z*-14)



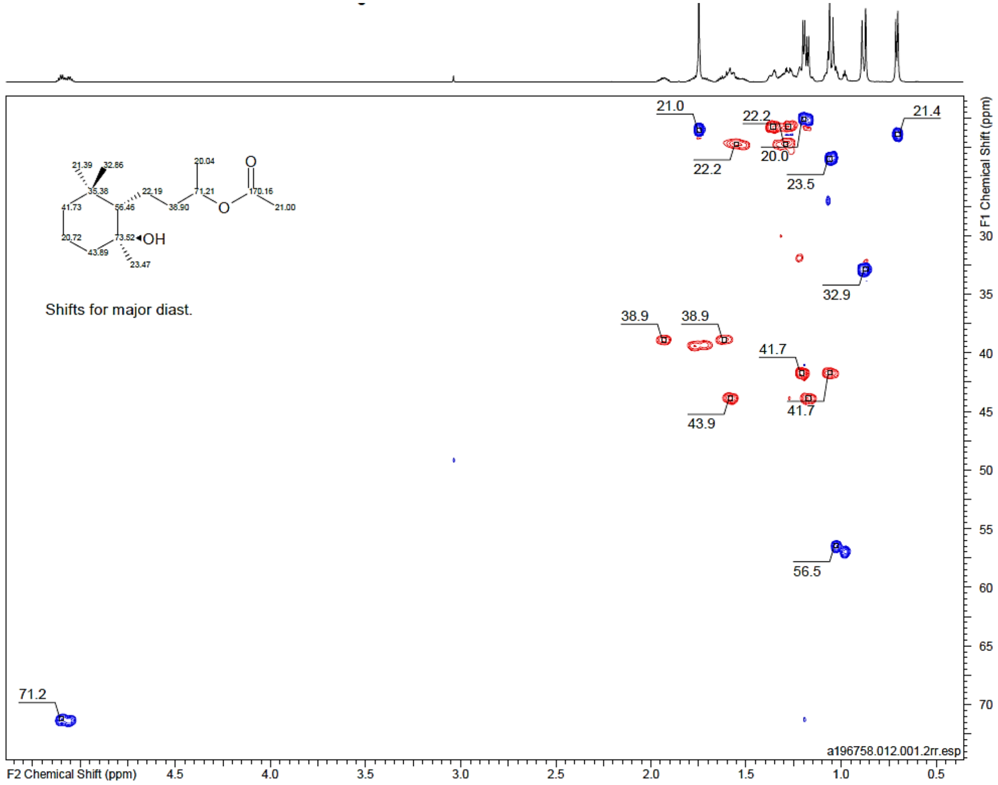


5. 4-((1S,2S)-2-hydroxy-2,6,6-trimethylcyclohexyl)butan-2-yl acetate (15)

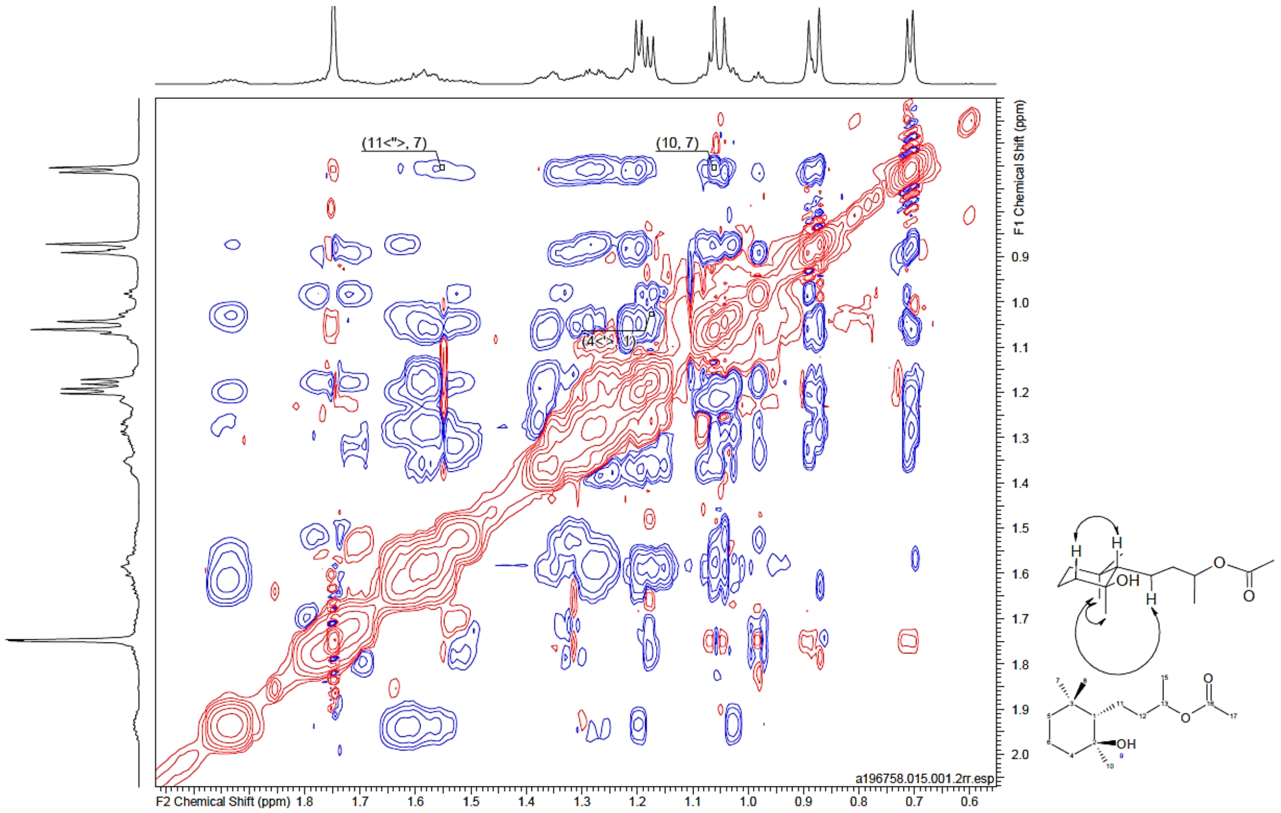




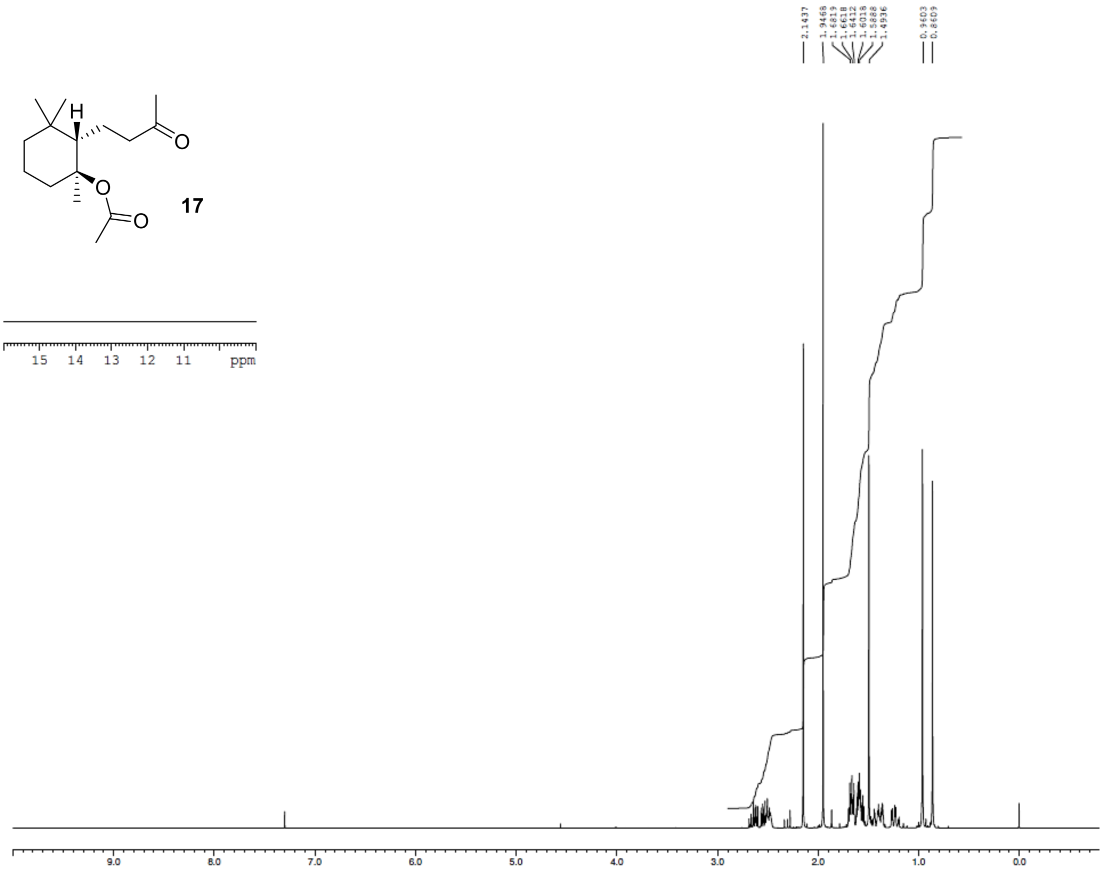
5.1. *HSQC of 4-((1S,2S)-2-hydroxy-2,6,6-trimethylcyclohexyl)butan-2-yl acetate (**15**)*

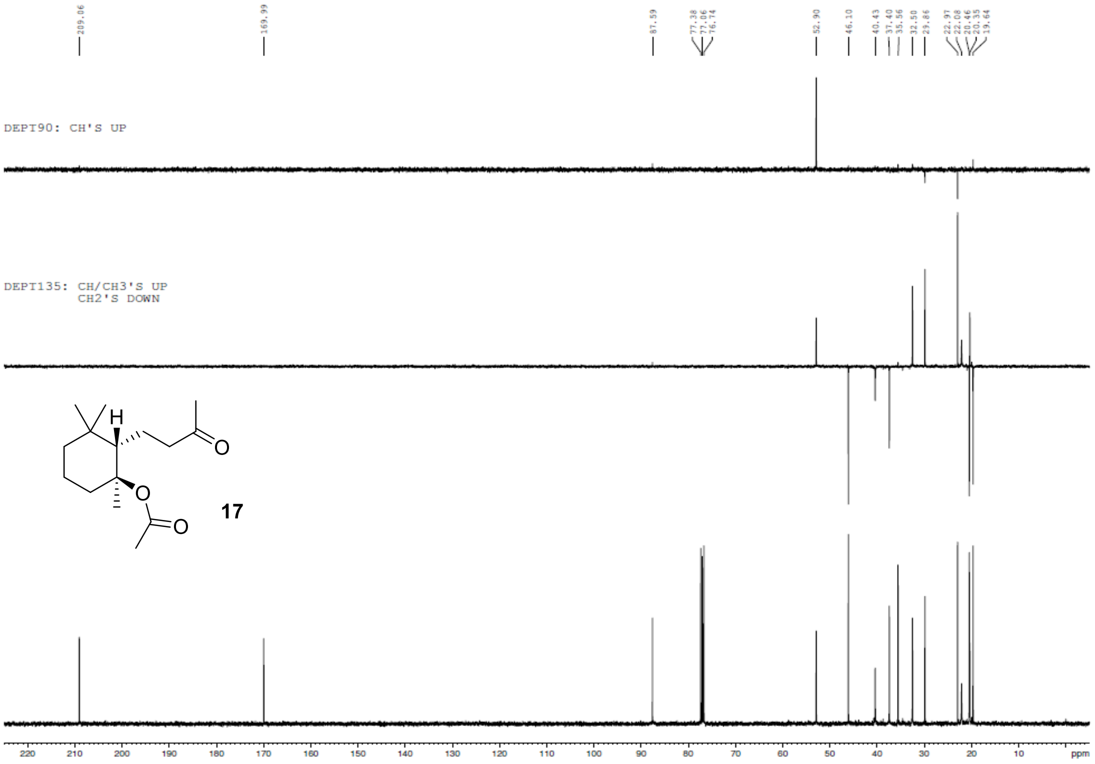


5.2. *NOESY of 4-((1S,2S)-2-hydroxy-2,6,6-trimethylcyclohexyl)butan-2-yl acetate (**15**)*

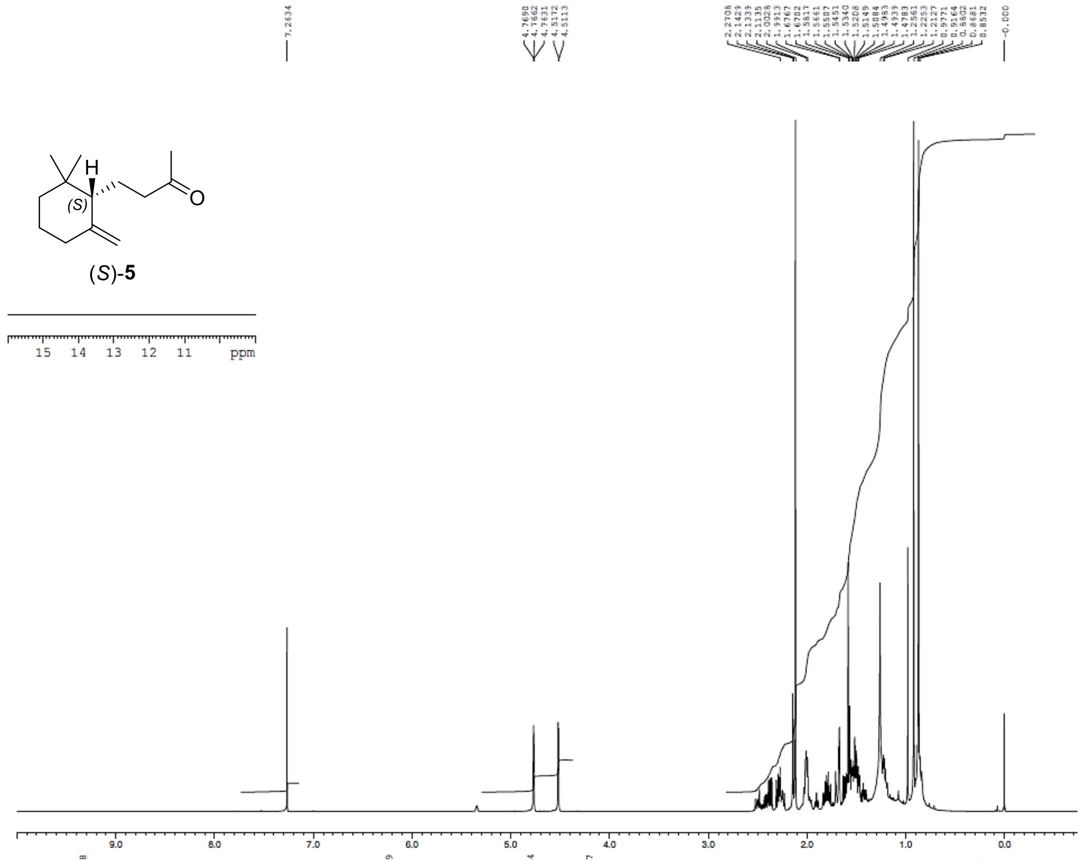


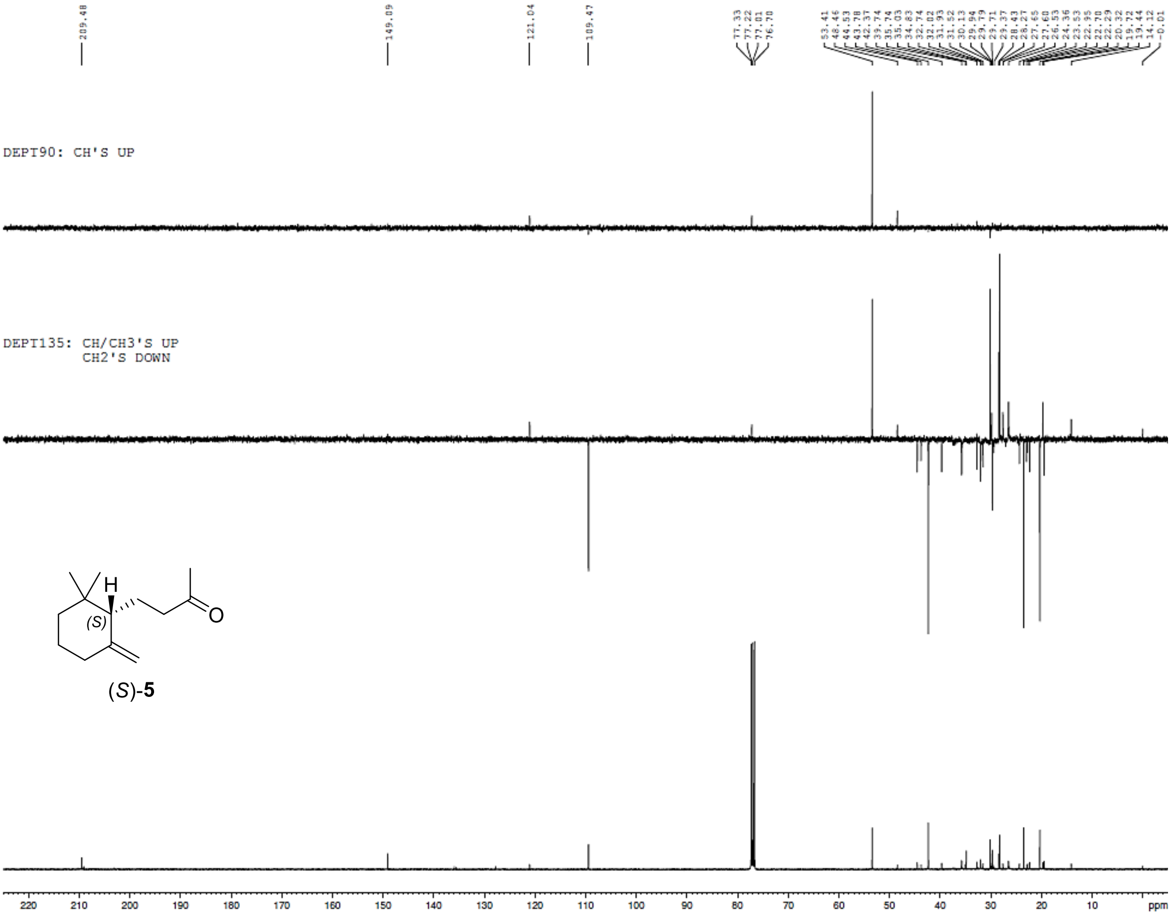
6. (1S,2S)-(+)-1,3,3-trimethyl-2-(3-oxobutyl)cyclohexyl acetate (17)



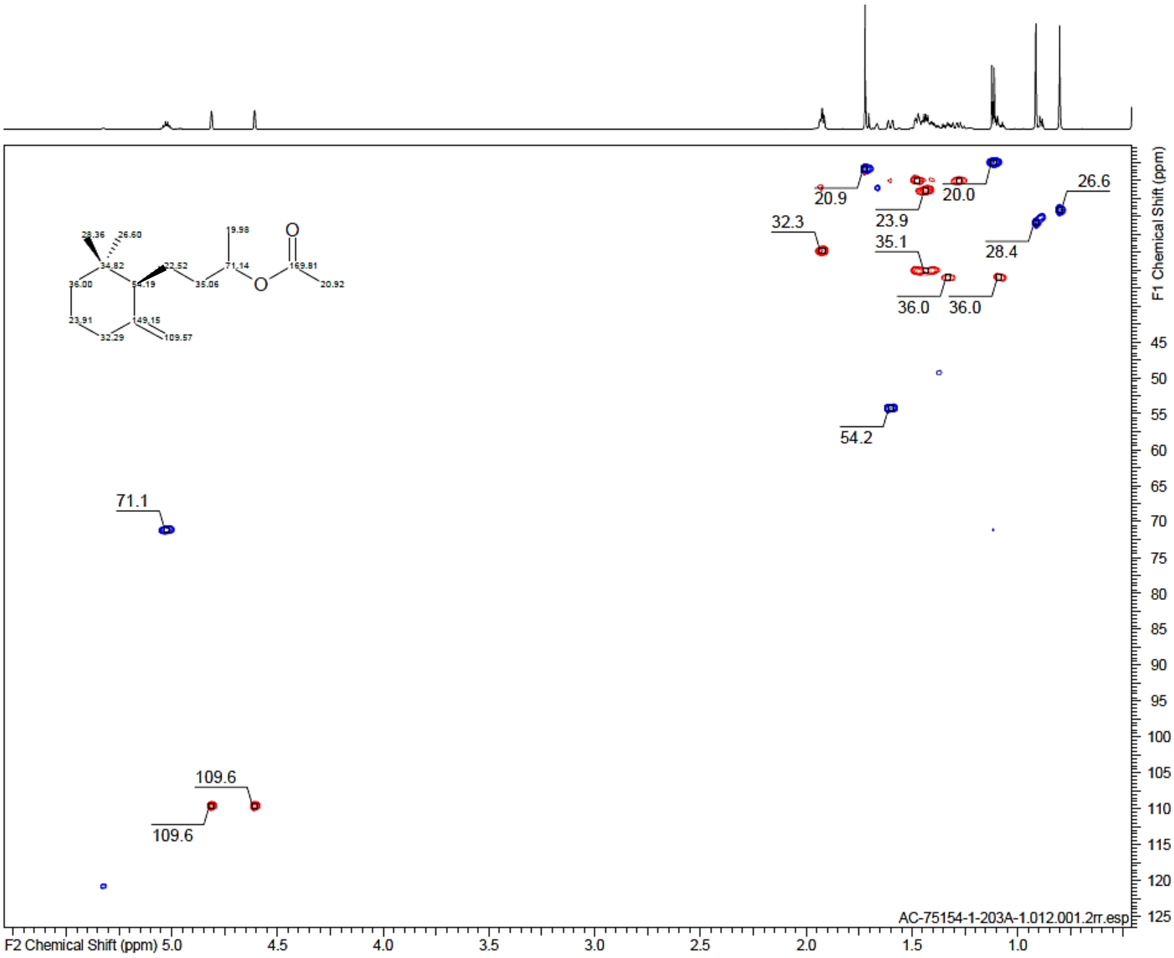


7. (*S*)-(+)-γ-dihydroionone ((S)-5) ex *E*-Tangerinol (γ:α:β 80:9:11)

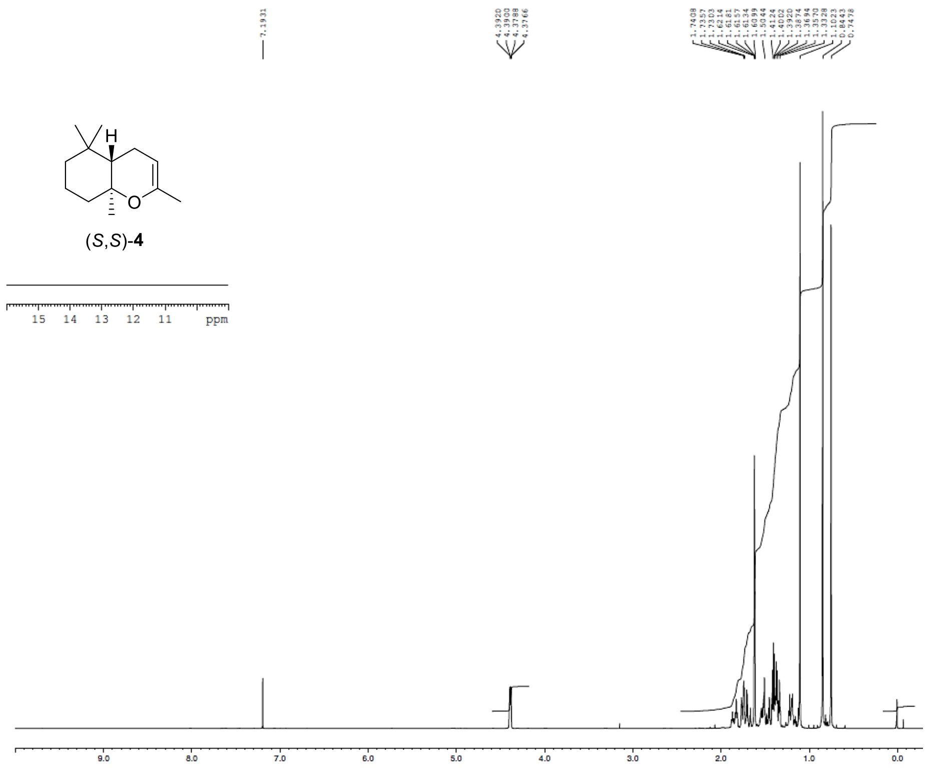


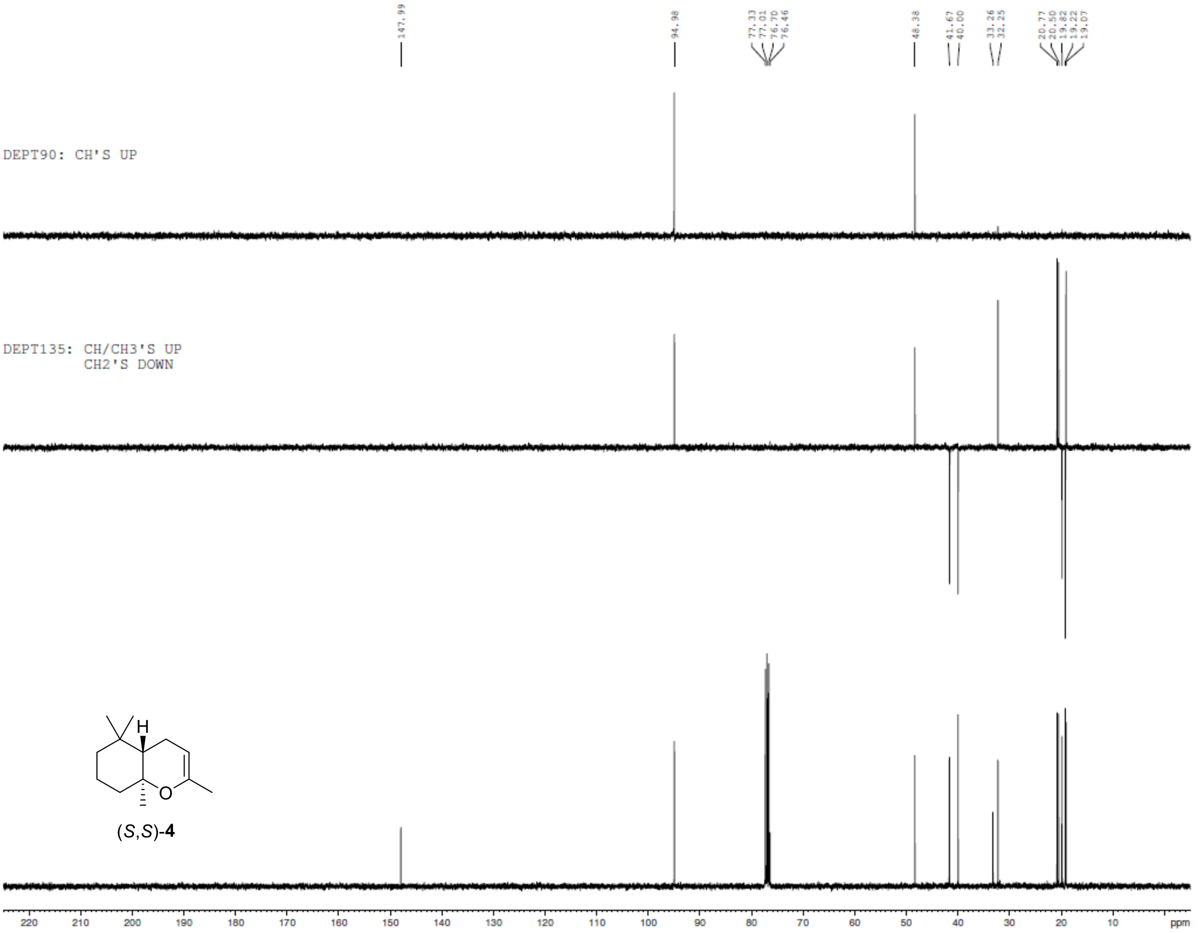


8. HSQC of 4-((*R*)-2,2-dimethyl-6-methylenecyclohexyl)butan-2-yl acetate (16)

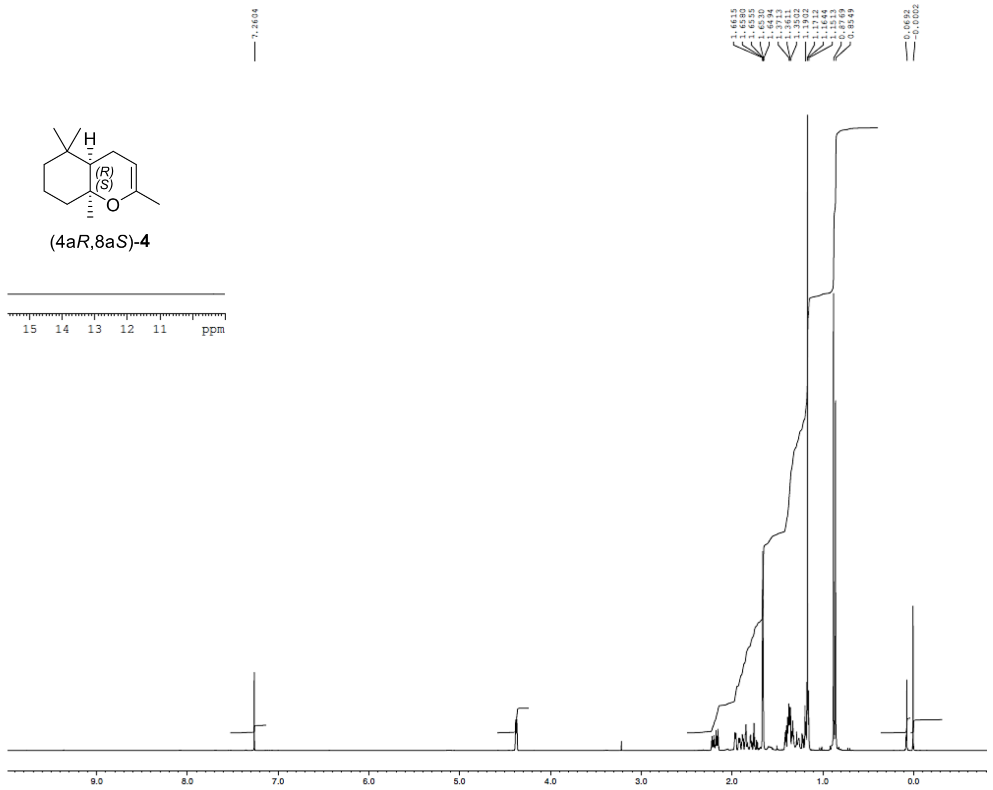


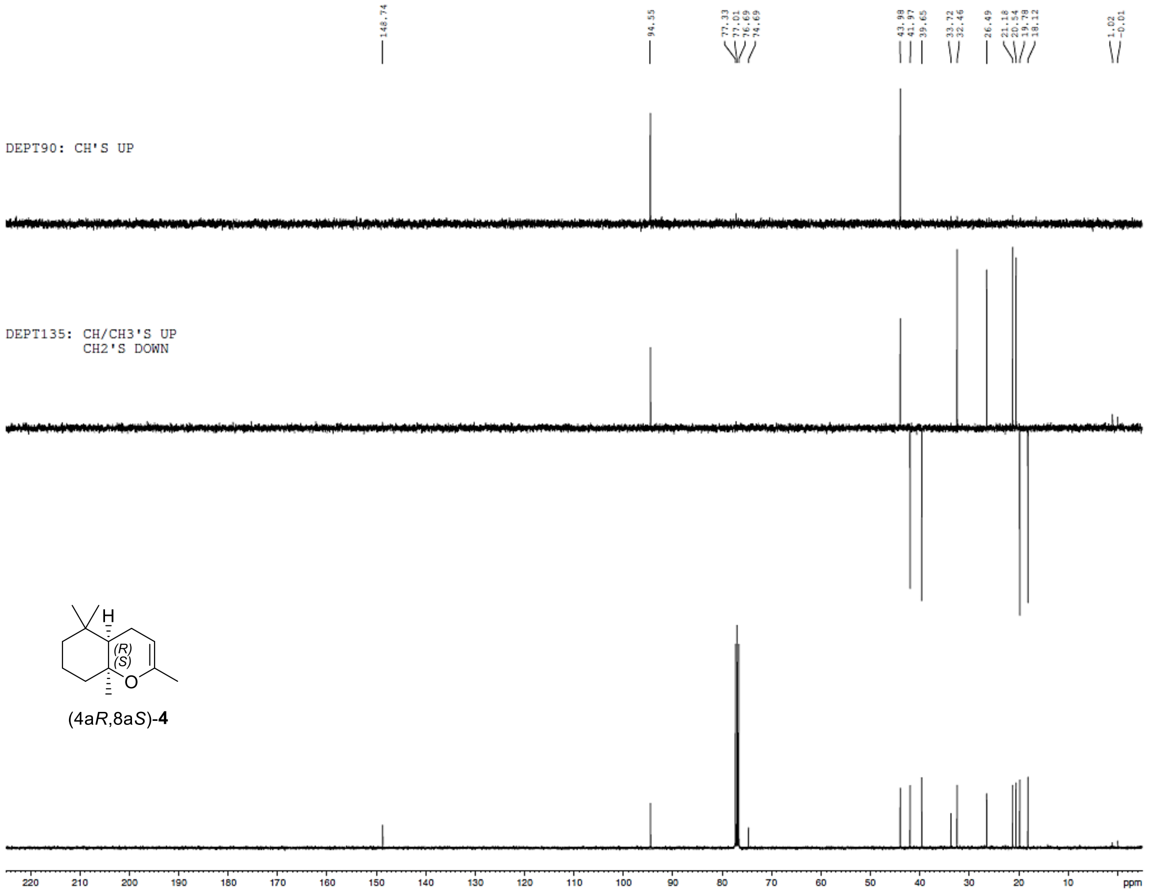
9. (4a*S*,8a*S*)-2,5,5,8a-tetramethyl-4a,5,6,7,8,8a-hexahydro-4H-chromene (*S*,*S*)-4





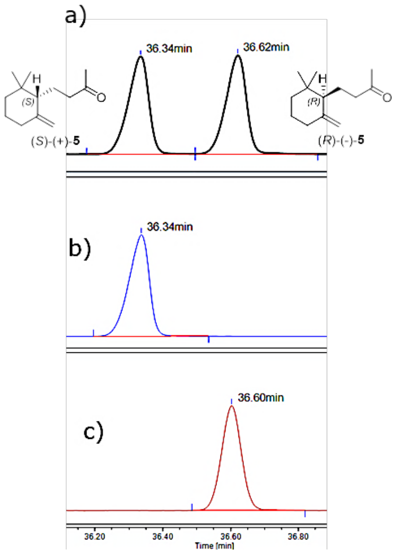
10. (4a*R*,8a*S*)-2,5,5,8a-tetramethyl-4a,5,6,7,8,8a-hexahydro-4H-chromene (*R*,*S*)-4





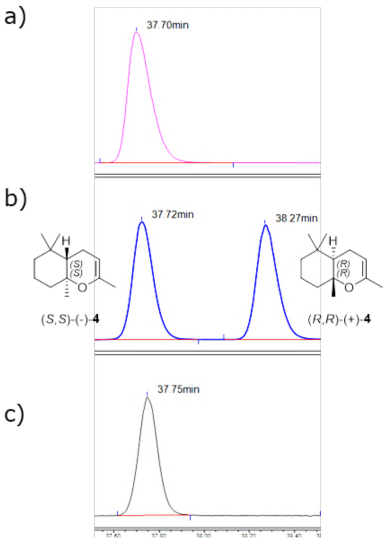
E. Chiral GC-analysis of SHC-cyclization products

1. (*R*)- and (*S*)- γ-dihydroionone (5)



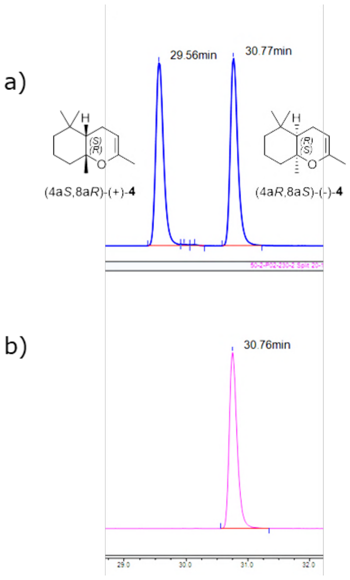
a) racemic γ-dihydroionone b) (*S*)-**5** ex *E*-tangerinol with *Aac*SHC 215G2, >99.9% e.e. c) (*R*)-**5** ex (*E/Z*)-geranylacetone (**2**) with AciSHC_R2.3, >99 % e.e.

2. trans bicyclic enolether (*S,S*)-4



a) (*S*,*S*)-**4** ex *E*-**2** by *Aac*SHC 215G2, >99.9% e.e. b) racemic (*S*,*S*)-**4** c) (*S*,*S*)-**4** ex (*E/Z*)-geranylacetone (**2**) by *Aci*SHC_R2.3, > 95% e.e.

3. cis bicyclic enolether (4a*R*,8a*S*-4)



a) racemic (4a*S*,8a*R*)-**4** b) (4a*R*,8a*S*)-(-)-**4** ex *Z*-**2** by *Aac*SHC 215G2, >99.9% e.e. (absolute configuration assigned tentatively)

F. Sequences

1. pET28b(+)

ctcgagcaccaccaccaccaccactgagatccggctgctaacaaagcccgaaaggaagctgagttggctgctgccaccgctgagcaataactagcataac
cccttggggcctctaaacgggtcttgaggggttttttgctgaaaggaggaactatatccggattggcgaatgggacgcgccctgtagcggcgcattaagcgcgg
cgggtgtggtggttacgcgcagcgtgaccgctacacttgccagcgccctagcgcccgctcctttcgctttcttcccttcctttctcgccacgttcgccggctttccccgt
caagctctaaatcgggggctccctttagggttccgatttagtgctttacggcacctcgaccccaaaaaacttgattagggtgatggttcacgtagtgggccatcgcc
ctgatagacggtttttcgccctttgacgttggagtccacgttctttaatagtggactcttgttccaaactggaacaacactcaaccctatctcggtctattcttttgatttat
aagggattttgccgatttcggcctattggttaaaaaatgagctgatttaacaaaaatttaacgcgaattttaacaaaatattaacgtttacaatttcaggtggcactttt
cggggaaatgtgcgcggaacccctatttgtttatttttctaaatacattcaaatatgtatccgctcatgaattaattcttagaaaaactcatcgagcatcaaatgaaac
tgcaatttattcatatcaggattatcaataccatatttttgaaaaagccgtttctgtaatgaaggagaaaactcaccgaggcagttccataggatggcaagatcctg
gtatcggtctgcgattccgactcgtccaacatcaatacaacctattaatttcccctcgtcaaaaataaggttatcaagtgagaaatcaccatgagtgacgactgaa
tccggtgagaatggcaaaagtttatgcatttctttccagacttgttcaacaggccagccattacgctcgtcatcaaaatcactcgcatcaaccaaaccgttattcatt
cgtgattgcgcctgagcgagacgaaatacgcgatcgctgttaaaaggacaattacaaacaggaatcgaatgcaaccggcgcaggaacactgccagcgcat
caacaatattttcacctgaatcaggatattcttctaatacctggaatgctgttttcccggggatcgcagtggtgagtaaccatgcatcatcaggagtacggataaaa
tgcttgatggtcggaagaggcataaattccgtcagccagtttagtctgaccatctcatctgtaacatcattggcaacgctacctttgccatgtttcagaaacaactct
ggcgcatcgggcttcccatacaatcgatagattgtcgcacctgattgcccgacattatcgcgagcccatttatacccatataaatcagcatccatgttggaatttaat
cgcggcctagagcaagacgtttcccgttgaatatggctcataacaccccttgtattactgtttatgtaagcagacagttttattgttcatgaccaaaatcccttaacgt
gagttttcgttccactgagcgtcagaccccgtagaaaagatcaaaggatcttcttgagatcctttttttctgcgcgtaatctgctgcttgcaaacaaaaaaaccacc
gctaccagcggtggtttgtttgccggatcaagagctaccaactctttttccgaaggtaactggcttcagcagagcgcagataccaaatactgtccttctagtgtagc
cgtagttaggccaccacttcaagaactctgtagcaccgcctacatacctcgctctgctaatcctgttaccagtggctgctgccagtggcgataagtcgtgtcttacc
gggttggactcaagacgatagttaccggataaggcgcagcggtcgggctgaacggggggttcgtgcacacagcccagcttggagcgaacgacctacaccg
aactgagatacctacagcgtgagctatgagaaagcgccacgcttcccgaagggagaaaggcggacaggtatccggtaagcggcagggtcggaacagga
gagcgcacgagggagcttccagggggaaacgcctggtatctttatagtcctgtcgggtttcgccacctctgacttgagcgtcgatttttgtgatgctcgtcaggggg
gcggagcctatggaaaaacgccagcaacgcggcctttttacggttcctggccttttgctggccttttgctcacatgttctttcctgcgttatcccctgattctgtggataa
ccgtattaccgcctttgagtgagctgataccgctcgccgcagccgaacgaccgagcgcagcgagtcagtgagcgaggaagcggaagagcgcctgatgcgg
tattttctccttacgcatctgtgcggtatttcacaccgcatatatggtgcactctcagtacaatctgctctgatgccgcatagttaagccagtatacactccgctatcgct
acgtgactgggtcatggctgcgccccgacacccgccaacacccgctgacgcgccctgacgggcttgtctgctcccggcatccgcttacagacaagctgtgac
cgtctccgggagctgcatgtgtcagaggttttcaccgtcatcaccgaaacgcgcgaggcagctgcggtaaagctcatcagcgtggtcgtgaagcgattcacag
atgtctgcctgttcatccgcgtccagctcgttgagtttctccagaagcgttaatgtctggcttctgataaagcgggccatgttaagggcggttttttcctgtttggtcactg
atgcctccgtgtaagggggatttctgttcatgggggtaatgataccgatgaaacgagagaggatgctcacgatacgggttactgatgatgaacatgcccggttac
tggaacgttgtgagggtaaacaactggcggtatggatgcggcgggaccagagaaaaatcactcagggtcaatgccagcgcttcgttaatacagatgtaggtgt
tccacagggtagccagcagcatcctgcgatgcagatccggaacataatggtgcagggcgctgacttccgcgtttccagactttacgaaacacggaaaccgaa
gaccattcatgttgttgctcaggtcgcagacgttttgcagcagcagtcgcttcacgttcgctcgcgtatcggtgattcattctgctaaccagtaaggcaaccccgcc
agcctagccgggtcctcaacgacaggagcacgatcatgcgcacccgtggggccgccatgccggcgataatggcctgcttctcgccgaaacgtttggtggcgg
gaccagtgacgaaggcttgagcgagggcgtgcaagattccgaataccgcaagcgacaggccgatcatcgtcgcgctccagcgaaagcggtcctcgccga
aaatgacccagagcgctgccggcacctgtcctacgagttgcatgataaagaagacagtcataagtgcggcgacgatagtcatgccccgcgcccaccggaa
ggagctgactgggttgaaggctctcaagggcatcggtcgagatcccggtgcctaatgagtgagctaacttacattaattgcgttgcgctcactgcccgctttccag
tcgggaaacctgtcgtgccagctgcattaatgaatcggccaacgcgcggggagaggcggtttgcgtattgggcgccagggtggtttttcttttcaccagtgagac
gggcaacagctgattgcccttcaccgcctggccctgagagagttgcagcaagcggtccacgctggtttgccccagcaggcgaaaatcctgtttgatggtggtta
acggcgggatataacatgagctgtcttcggtatcgtcgtatcccactaccgagatatccgcaccaacgcgcagcccggactcggtaatggcgcgcattgcgcc
cagcgccatctgatcgttggcaaccagcatcgcagtgggaacgatgccctcattcagcatttgcatggtttgttgaaaaccggacatggcactccagtcgccttc
ccgttccgctatcggctgaatttgattgcgagtgagatatttatgccagccagccagacgcagacgcgccgagacagaacttaatgggcccgctaacagcgcg
atttgctggtgacccaatgcgaccagatgctccacgcccagtcgcgtaccgtcttcatgggagaaaataatactgttgatgggtgtctggtcagagacatcaaga
aataacgccggaacattagtgcaggcagcttccacagcaatggcatcctggtcatccagcggatagttaatgatcagcccactgacgcgttgcgcgagaaga
ttgtgcaccgccgctttacaggcttcgacgccgcttcgttctaccatcgacaccaccacgctggcacccagttgatcggcgcgagatttaatcgccgcgacaattt
gcgacggcgcgtgcagggccagactggaggtggcaacgccaatcagcaacgactgtttgcccgccagttgttgtgccacgcggttgggaatgtaattcagctc
cgccatcgccgcttccactttttcccgcgttttcgcagaaacgtggctggcctggttcaccacgcgggaaacggtctgataagagacaccggcatactctgcgac
atcgtataacgttactggtttcacattcaccaccctgaattgactctcttccgggcgctatcatgccataccgcgaaaggttttgcgccattcgatggtgtccgggatc
tcgacgctctcccttatgcgactcctgcattaggaagcagcccagtagtaggttgaggccgttgagcaccgccgccgcaaggaatggtgcatgcaaggagatg
gcgcccaacagtcccccggccacggggcctgccaccatacccacgccgaaacaagcgctcatgagcccgaagtggcgagcccgatcttccccatcggtg
atgtcggcgatataggcgccagcaaccgcacctgtggcgccggtgatgccggccacgatgcgtccggcgtagaggatcgagatctcgatcccgcgaaatta
atacgactcactataggggaattgtgagcggataacaattcccctctagaaataattttgtttaactttaagaaggagatata

2. *Bth*SHC

atgggcttattatacgaaaaagcgcatgaagaaatagcgagaagaacaactgcacttcaaacaatgcaacggcaagatggtacgtggcagttttgttttgaaggagcgccgctaacagatt
gtcatatgattttttattaaaattattaggtagagataaagagatagaaccgtttgtaaaaagattagcatcacttcaaacaaatgaaggaacatggaaattgtatgaagatgaaatgggtggta
atttatctgctacaattcaatcttatgctgccttacttgcatcagaaaaatatacaaaagaagatgcgaatatgaagcgagcggaaatgtttataaatgagcgcggtggggtagcgcgtgctcat
tttatgacgaagtttttattagcgattcatggagaatatgaatatccttctctctttcatttgccaacaccaattatgtttctgcagaatgattcccctctcagtatatttgaattgagtagctcagcacgtat
ccatttaattccgatgatgttgtgtttaaataaaagatttcgagtagggaaaaagttattgccaaatttaaatcatattgcaggcgggggcggagaatggtttcgggaggatcggtctccagtttttc
aaacgttagtaagtgacgtgaagaaaattataacgtatccactttctttgcatcataaaggatatgaggaagtagaacgttttatgaaagagcgtattgatgaaaatggaacattatatagttac
gcaactgcctcgttttatatgatttatgctttacttgcattaggacattctattcactcaccaattattcagaaggctataatgggaatcacatcttatatatggaagatggagagagggagccatttg
caaaactctccgtcaactatatgggatacagctttactcagttatgctttgcaagaagctcaagttccgaaagcaagtaaagtgattcaaaatgcatcagcgtatttactaagaaaacagcaaa
caaagaaagtagattggagtgtacatgcaccggatctcttcccaggtggttggggcttttcggatgtgaatacgacgattccagatattgatgatacaactgctgcgttaagagcattggcgcg
aagtagagggaacgaaaatgtagacaatgcttggaagcgagcggttaattgggttaaaggattgcaaaataatgatggtggttggggggcttttgaaaaaggggtaacgagccgtatatta
gcaaatttaccaatcgaaaatgcaagtgatatgattacagatccttctacaccagatattacaggaagagtgttagaatttttcgggacgtatacgcaaaatgaattgcccgagaaacaaaaa
caaagtgcgataaattggttaacgaatgcacaagaggaaaatggatcatggtatgggaaatggggggatttgttatatatatggtacgtgggcggttatgactggtttacggtcactaggaactc
catctagcaacccatcattaaaacgagccgctttatggcttgaacatatacagcatgaagatggtggctggggagaatcttgccacagtagtgtggagaaaaggttcgttactttaccatttagt
acaccatcccaaacagcatgggcgttagatgccctcatttcttactatgataaagaaacgccagtcattcgcaaaggtatttcatatttgctctccaacccttatgtaaatgaaaaatatcctactg
gaacaggcttgccaggtgggttttatattcgttatcatagttatgctcatatatatccgttgcttactttggcacattatttaaaaaaatatagaaaataa

3. *Aac*SHC

atggcagaacagctggttgaagcccctgcctatgcacgtaccctggatcgtgcagttgaatatctgctgagctgtcagaaagatgaaggttattggtggggtccgctgctgagcaatgttacaa
tggaagcagaatatgtgctgctgtgtcatattctggatcgcgttgatcgtgatcgcatggaaaaaattcgtcgttatctgctgcatgaacagcgtgaagatggcacctgggcactgtatcctggtg
gtccgcctgatctggataccaccattgaagcctatgttgcactgaaatatatcggtatgagccgtgatgaagaaccgatgcagaaagcactgcgttttattcagagccaaggtggtattgaaag
cagccgtgtttttacccgtatgtggctggcactggttggtgaatatccgtgggaaaaagttccgatggttccgcctgaaattatgtttctgggtaaacgtatgccgctgaacatttatgaatttggtag
ctgggcacgtgcaaccgttgttgccctgagcattgtgatgagccgtcagccggtttttccgctgccggaacgtgcccgtgtgccggaactgtatgaaaccgatgttcctccgcgtcgtcgtggtg
caaaaggtggtggtggttggatttttgatgcactggaccgtgcactgcatggttatcagaaactgagcgttcatccgtttcgtcgtgcagcagaaattcgtgcactggattggctgctggaacgtc
aagccggtgatggtagttggggtggtattcagcctccgtggttttatgcactgattgccctgaaaattctggatatgacccagcatccggcatttatcaaaggttgggaaggtctggaactgtacg
gtgttgaactggattatggtggctggatgtttcaggcaagcattagtccggtttgggataccggtctggcagttctggcactgcgtgcagccggtctgcctgcagatcatgaccgtctggttaaag
caggcgaatggctgttagatcgtcagattaccgttcctggtgattgggcagttaaacgtccgaatctgaaacctggtggttttgcatttcagttcgacaatgtttattatccggatgtggatgataccg
cagttgttgtttgggcactgaataccctgcgtctgcctgatgaacgtcgtcgccgtgatgcaatgaccaaaggttttcgttggattgttggtatgcagagcagcaatggcggttggggtgcctatga
tgttgataataccagcgatctgccgaaccatattccgttttgtgattttggtgaagttaccgatccgcctagcgaagatgttaccgcacatgttctggaatgttttggcagctttggttatgatgatgcct
ggaaagttattcgtcgcgctgtggaatatctgaaacgtgaacagaaaccggatggttcatggtttggtcgctggggtgttaattatctgtatggtacaggtgcagttgttagcgcactgaaagcag
ttggtattgatacccgtgaaccgtatattcagaaagccctggattgggttgaacagcatcagaatccggacggtggctggggtgaagattgtcgtagctatgaagatcctgcgtatgcaggtaa
aggtgcaagcaccccgagccagaccgcatgggctctgatggccctgattgccggtggtcgtgccgaaagcgaagcagcacgtcgcggtgttcagtatctggttgaaacccagcgtccgga
tggcggatgggatgaaccttattacaccggcaccggttttccgggtgatttttatctgggttataccatgtatcgtcatgtgtttccgacactggccctgggtcgttataaacaggcaattgaacgtc
gctaa

4. *Bja*SHC1

atggatagcgttaatgcaaccgcacgtgaagcaaaagaaagcaaaattagcgaaagcgaaattctggaaagcagcattgcaagcgcaacccagggtgttctgggttttcagcagagtga
tggtcattgggtttttgaactggaagcagattgtaccattccggcagaatatgttctgctgcgtcattatctggcagaaccggttgataccgttctggaagcaaaaattggtaattatctgcgtcgtgt
tcagggtgcacatggtggttggcctctggttcatgatggtgaatttgatatgagcgcaagcgtgaaagcatattttgccctgaaaatgattggcgatagcattgatgctccgcacatggttcgtgcc
cgtgaagccattcatgcacgtggtggtgcaattcatagcaatgtttttacccgttttatgctggccatgtttggtattgttacctggcgtgcagttccggtgctgccgattgaaattatgctgctgccgttt
tggagcccgtttcatattaacaaaatcagctattgggcacgtaccacaatggttccgctgatggttattgcagcactgaaaccgcgtgcaaaaaatccgaaaggtgttggtattgatgaactgttt
ctgcaagatccgcgtagcattggtatgaccgcaaaagcaccgcatcagagcatggcatggtttctgctgtttcgtagcctggatgcaattctgcgtgttattgaaccgctgtttccgaaaagcctg
cgtaaacgtgcaattgataccgcactggcatttagcgaagaacgtctgaatggtgaagatggtatgggtgcaatttatccgcctatggcaaatctggtgatgatgtatgatgcactgggcaaag
atgaaaattatccgccacgtgcagttacccgtcgcggtatcgataaactgctggttattggagatgatgaagcatattgtcagccgtgtgttagtccggtttgggataccacactgaccgcacat
gcactgctggaagccggtggtgataaagcaggtccggcagcaaaacatggtctggattggctgattccgaaacaagagctggaagttaaaggtgattgggcagttaaacgtccggatgttc
gtccaggtggttgggcatttcagtataataacgcatattatccggatctggatgatacagcagttgttgttatgagcatggatcgtatgcgtcgtgaacatggtgttaccggttatgatagcgcaatt
gatcgtggtcgtgaatggattgaaggtatgcagtcagatgatggtggctgggcagcatttgatgttaataatctggaatattacctgaacaacatcccgtttagcgatcatggtgccctgctggac
cctccgaccgaagatgttaccgcacgttgtgttagcatgctggcacagctgggtgaaaccgcaaaaaccagcaaacatgttgcagatggtgttgcatatctgcgtaaaacccagcatccgg
aaggtagctggtatggtcgttggggtatgaattttatctatggcacctggtcagttctgtgtgcactgaatatggcaggcgttcgtcatgatgatccgatgattcgtaaagcagcagattggctggc
aagcattcagaataaagatggcggttggggtgaagataccgttagctatcgtctggattataaaggttgggaagcagcaccgagcaccgcaagccagaccgcatgggctctgctggcact
gatggcagcgggtgaagttgatcatccggcagttgcccgtggtgtggaatatctgattgcaacccagaatgaaaaaggtctgtgggatgaacagcgttataccgcaaccggttttccgcgtgtt
ttttacctgcgttatcatggctatagcaaatttttcccgctgtggggtttagcacgttatcgtaatctgcgcaataccaatagccgtgttgttggtgtgggtatgtaa

5. *Zmo*SHC1

atgggtattgatcgtatgaatagcctgagccgtctgctgatgaaaaaaatctttggtgcagagaaaaccagctataaaccggcaagcgataccattattggtacagataccctgaaacgtccg
aatcgtcgtccggaaccgaccgcaaaagttgataaaaccatctttaaaacgatgggcaacagcctgaataacaccctggttagcgcatgtgattggctgattggtcagcagaaaccggatg
gtcattgggttggtgcagttgaaagcaatgcaagcatggaagcagaatggtgtctggcactgtggtttctgggtttagaagatcatccgctgcgtccgcgtctgggtaatgcactgctggaaatg
cagcgtgaagatggtagctggggtgtttattttggtgccggtaatggtgatattaatgcaaccgttgaagcctatgcagcactgcgtagcctgggttatagcgcagataatccggttctgaaaaa
agcagcagcatggattgcagaaaaaggtggtctgaaaaacattcgtgtgtttacccgttattggctggcactgattggtgaatggccgtgggaaaaaaccccgaatctgcctccggaaattat
ctggtttccggataattttgtgttcagcatctataactttgcacagtgggcacgtgcaacaatggttccgattgcaattctgagcgcacgtcgtccgagccgtccgttacgtccgcaggatcgtctg
gatgaactgtttccggaaggtcgtgcacgttttgattatgaactgccgaaaaaagaaggcatcgatctttggagccagtttttcgtaccaccgatcgtggtctgcattgggttcagagcaatctgc
tgaaacgtaatagcctgcgtgaagcagcaattcgtcatgttctggaatggattattcgtcatcaggatgcagatggtggttggggtggtattcagcctccgtgggtttatggtctgatggccctgca
tggtgaaggttatcagctgtatcatccggttatggcaaaagcactgagtgcactggatgatcctggttggcgtcatgatcgtggtgaaagcagctggattcaggcaaccaatagtccggtttgg
gataccatgctggccctgatggcactgaaagatgcaaaagcagaagatcgttttacaccggaaatggataaagcagccgactggctgctggcacgtcaggttaaagttaaaggtgattgg
agcattaaactgccggatgttgaacctggtggctgggcatttgaatatgccaatgatcgttatcctgataccgatgataccgcagttgcgctgattgcactgagcagctatcgtgataaagaaga
gtggcagaaaaaaggcgttgaagatgcaattacccgtggtgttaattggttaattgcaatgcagagcgaatgtggcggttggggagcatttgataaagataataatcgtagcatcctgagcaa
aatcccgttttgtgattttggcgaaagcattgatccgcctagcgttgatgttaccgctcatgtgctggaagcatttggcaccctgggtctgagccgtgatatgccggttattcagaaagcaattgatt
atgtgcgtagcgaacaagaggcagaaggtgcctggtttggccgttggggtgtgaattatatctatggtacaggtgcagttctgcctgcactggcagcaattggtgaagatatgacccagccgt
atattaccaaagcctgcgattggctggttgcacatcagcaagaggatggcggatggggtgaaagctgtagcagctatatggaaattgatagcattggtaaaggtccgaccacaccgagcca
gaccgcatgggcactgatgggtctgattgcagcaaaccgtccggaagattatgaagcaattgcaaaaggttgccactatctgattgatcgtcaagaacaggatggttcctggaaagaagaa
gaattcaccggcaccggttttccaggttatggtgttggtcagacaattaaactggacgatccggcactgagtaaacgtctgctgcagggtgcagaactgagtcgtgcatttatgctgcgttatgat
ttttatcgtcagtttttcccgattatggcactgtcacgtgcagaacgtctgatcgatctgaacaattaa

6. *Bme*TC

atgatcatcctgctgaaagaagtgcagctggaaattcagcgtcgtattgcatatctgcgtccgacacagaaaaatgatggtagctttcgttattgctttgagacaggtgttatgccggatgcatttct
gattatgctgctgcgtacctttgatctggataaagaagttctgattaaacagctgaccgaacgtattgttagcctgcagaatgaagatggtctgtggaccctgtttgatgatgaagaacataatctg
agcgcaaccattcaggcatataccgcactgctgtatagcggttattatcagaaaaacgatcgcattctgcgtaaagccgaacgctatattatcgatagcggtggtattagccgtgcacattttctg
acccgttggatgctgagcgttaatggtctgtatgaatggccgaaactgttttatctgccgctgagcctgctgctggttccgacctatgttccgctgaacttttatgaactgagcacctatgcacgcatt
cattttgttccgatgatggttgcaggcaacaaaaaattcagcctgaccagccgtcataccccgagcctgagccatctggatgttcgtgaacagaaacaagaaagcgaagaaacaacccaa
gaaagccgtgcaagtatttttctggttgatcatctgaaacagctggcaagcctgccgagctatattcataaactgggttatcaggcagcggaacgttatatgctggaacgcattgaaaaagatg
gcaccctgtatagctatgcaaccagcaccttttttatgatttatggtctgctggcactgggctacaaaaaagatagctttgtgattcagaaagccattgatggtatttgtagcctgctgagtacctgta
gcggtcatgttcatgttgaaaatagcaccagcaccgtttgggatacagccctgctgagctatgcactgcaagaagcaggcgttccgcagcaggacccgatgattaaaggtacaacccgttat
ctgaaaaaacgccagcataccaaattaggcgattggcagtttcataatccgaataccgcacctggtggttggggttttagcgatattaacaccaataatccggatctggatgataccagcgca
gcaattcgtgcactgagccgtcgtgcacagaccgataccgattatctggaaagctggcagcgtggtattaattggctgctgtcaatgcagaacaaagatggtggttttgcagcctttgaaaaaa
acaccgatagcatcctgtttacctatctgcctctggaaaatgcaaaagatgcagcaaccgatccggcaaccgcagatctgaccggtcgtgttctggaatgtctgggtaattttgcaggtatgaat
aaaagccatccgagcattaaagcagcagtgaaatggctgtttgatcatcagctggataatggtagctggtatggtcgttggggtgtttgttatatctatggcacctgggcagcaattaccggtctg
cgtgcagttggtgttagcgcaagcgatccgcgtattatcaaagccattaactggctgaaaagcatccagcaagaggatggtggctttggtgaaagctgttatagcgccagcctgaaaaaata
cgtgccgctgtcatttagcaccccgagtcagaccgcatgggcattagatgcactgatgaccatttgtccgctgaaagatcgtagcgttgaaaaaggtatcaagtttctgctgaatccgaatctga
cagaacagcagacccattatccgaccggtattggtctgcctggtcagttttatatccagtatcacagctataacgatatctttccgctgctggccctggcacattatgcaaaaaaacatagcagct
aa

7. *Gni*PNT

atgctgccgtataatcagaacagctataaagaagcactgcacggcggtcatgcagcacataatccgcctacactggaagaagcaattaaacgtagccaagaatttctgctggcacatcag
catccggaaggtttttggtggggtgatctggaatgtaatgttaccagcgcaagtcatacccctgatcctgtataaaatcctgggtattgcagatcgttatccgctgcacaaatttgagaaatatctgc
gtcgtatgcagtgtagtcatggtggttgggaaatgagctttggtgatggtggttatctgagcgcaaccattgaagcatatatttgtctgcgtctgctgaatgttccgcagagcgatccggcactgca
gcgtgcactgaaaaacattctggcacgtggtggtgttaccaaagcacgtgtttttaccaaagtttgtctggcactgttaggtggttttgattgggcagcactgccgagcctgcctccgtggctgatg
ctgtttccggcatggtttccgtggaacatttatgaagcagcaagctgggcacgcggttgtgttgttccgctgattgttctgctggaaaaaaaaccggtgtttcaggttaaaccggaagtcagctttg
atgaactgtatgttgaaggtcgtgcacatgcctgtaaagccctgccgtttagcgcacatgattgggttagcaacatttttgttgcagcagatcgtgcctttaaactgatggaacgttttggtgcagttc
cgtttcgtcagtggtcaattaaagaagccaaaaaatgggtgcttgatcgccaagaagaaatgggcgattttattggttataatcctccgatgctgtattttgccgtttgtctgaaactgtggggttatg
aagttaccgatccgctgttacagcgtgccctgctggcccataaaaaactgaccgttgaaaccgaagatgaatgttggctgcagagcagccagagtccggtttgggataccgcactggttattc
cagcactggttgaaagcggtctgcctccggatcatcctgcgctgcagaaagcaggtcagtggctgttagaaaaacaaattctgaaacatggcgactgggctctgaaaacaggtggtggtcg
catgcaggatgacattggtggtggctgggcatttcagtttgttaatagctggtatccggatgtggatgatagcgcagcagttgttattgcactgaactgcattaaaatgccggatgaggatgttaa
aaatggtgcaattgcccgttgcctgaaatggattgcatttatgcagggtcgtaatggcggatgggcagcatttgatcgtgatagcaatcagcgttggatggatgcaaccccgtttagtgatattga
agcaatgctggatgttagcaccgcagatgttaccgcacgtgttctggaaatggttggtctgatgcgtctgaaacacgcagcacagcctgcaaataattcactgggtaaagcacatcgtcatatt
agcaccgaaagcattgcccgtggtgttgattatctgaccaaagaacaagaaaaagaaggctgttggtggggacgttggggtgtgaattatatctatggcacccgtggtgcactgatgggtctg
agccaggttgcagcaaaaacacataaaaaagaaattgcgcgtggtgcagcatggctggttaaagttcagaacaaaaagaacgagaagaaacagggtgcacaggatggcggttgggg
tgaagcatgttttagctatgatgatcctgcaaccaaaggtcagaatagccgtagcaccgccagccagaccggttgggcaatgcagggactgctggcagccggtgaagttctgggtcgtaaa
tatgaaatggaagcagttgaagaaggtgtgcagtttctgttagatacccagcgtaaagatggtagctggtctgaagcagaatttaccggtggtggttttccgaaacactattatctgaaatacca
ctattttgcccagcattttccgctgagcgcactggcacgttatcgtgcccgtctgctgcagctgagccgtccgaaaaatcaggcataa

8. *Aci*SHC

atgacccaggcaagcgttcgtgaagatgcaaaagcagcactggatcgtgcagttgattatctgctgagcctgcaggatgaaaaaggtttttggaaaggtgaactggaaaccaacgttacca
ttgaagccgaagatctgctgctgcgtgaatttctgggtattcgtacaccggatattaccgcagaaaccgcacgttggattcgtgcaaaacagcgtagtgatggcacctgggcaaccttttatgat
ggtccgcctgatctgagcaccagcgttgaagcctatgttgcactgaaactggctggtgatgatccggcagcaccgcacatggaaaaagcagccgcatatattcgtggtgccggtggtgttga
acgtacccgtgtttttacccgtctgtggctggcactgtttggtctgtggccgtgggatgatctgccgacactgcctccggaaatgatttttctgccgagctggtttccgctgaacatttatgattggggtt
gttgggcacgtcagaccgttgtgccgctgaccattgttagcgcactgcgtccggttcgtccgattccgctgagtattgatgaaattcgtacaggtgcaccgcctccgcctcgcgatccggcatgg
accattcgtggtttttttcagcgtctggatgatttactgcgtggttatcgtcgtgttgcagatcatggtccggcacgtctgtttcgtcgtctggcaatgcgtcgtgcagcagaatggattattgcacgtca
agaggcagatggtagctggggtggtattcagcctccgtgggtttatagcctgattgcactgcatctgctgggttatccgctggatcatccggttctgcgtcgtggtctggatggtctgaatggttttac
cattcgcgaagaaacagcagatggtgcagttcgtcgcctggaagcatgtcagagtccggtttgggataccgcactggcagttacagcactgcgtgatgcaggtctgcctgccgatcatccgc
gtgttcaggcagcagcccgttggctggttggtgaagaggtgcgtgttgccggtgattgggcagtgcgtcgtccgggtctgcctcctggtggttgggcatttgaatttgccaatgataattacccgg
ataccgatgatacagcagaagttgttctggccctgcgtcgcgttcgtctggaagatgcagatcagcaggcgctggaagcagccgttcgtcgtgcaaccacctgggttattggtatgcagagca
ccgatggcggttggggtgcatttgatgcagataatacccgtgaactggtgctgcgtctgccgttttgtgattttggtgccgttattgatccgcctagcgcagatgttaccgcacatattgttgaaatgc
tggcagccctgggtatgcgtgatcatcctgccaccgttgcgggtgttcgttggctgctggcacatcaagaacctgatggtagttggtttggtcgttggggagcaaatcatatttatggtacgggtgc
agttgttccggcactgattgcagccggtgttagtccggatacaccgcctattcgtcgcgcaattcgctggctggaagaacatcagaatccggatggtggatggggtgaagatttacgtagctat
accgatcctgcactgtgggttggtcgtggtgttagcaccgcaagccagaccgcatgggcactgctggcattactggcagcgggtgaagaagcaagtccggcagttgatcgtggcgttcggtg
gctggttaccacacagcagcctgatggtggctgggatgagccgcattacaccggtacaggttttccgggtgatttctatattaactatcatctgtatcgcctggtgtttccgattagtgcactgggtc
gttatgttaatcgt

9. *Gvu*SHC

atggttgcagatgaacgtagcgttctgattgatgcactgaaacagagccaggcagcagatggtagctggcgttttccgtttgaaaccggtattagcaccgatgcctatatgattattctgctgcgt
accctggaaattaatgatgaaccgctgattcaggcactggttgaacgtattgaaagccgtcaagaggcaaatggtgcatggaaactgtttgccgatgaaggtgatggtaatgttaccgcaac
caccgaagcatactatgcactgctgtatagcggttatcgtaaaaaaaccgatccgcacatgcagaaagcaaaaatgcgtattctggaaatgggtggtctggaaaatgttcacctgtttaccaa
agttatgctggcactgacaggtcagcatccgtggcctcgtcgttttccgctgccgctggcatttttctgctgccaccgagctttccgctgaatatgtatgatctgagcgtttatggtcgtgccaatatg
attccgctgctggttgcagccgaacgtcgttatagccgtaaaaccgcaaaaagtccggatctgagtgatctggcagcaagccgtggtggttggcgtctgccggaaaatcgtgcactgtggtca
tatatcaaacgtgcactgaccggttttccggatgaactgcatgatgcagcaaaacagcgtgcagttcgttatatgtttggtcatattgaaccggatggcaccctgtatagctattttagcagcacct
ttctgtttatctttgccctgctggccctgggttatccgaaagatgatctgcatattgcacgtgccgttcgtggtctgcgtagcctgcgtaccgaaattgatggtcatacccacatgcagtataccaccg
caagcgtttggaataccgcactggcaagctatgccctgcaagaagcaggcgttccgagcaccgatcgtaccattgaaaaagcaaatcgttatctgctgagccgtcagcatattcgttatggtg
attgggcagttcataatccgcatagcctgcctggtggctggggttttagtgatgttaataccatgaatccggatgtggatgataccacagcagcactgcgtgcaattcgtcgtgcagcagcaaa
agaaaccgcatttcgtcatgcatgggatcgtgcgaatcgttggctgtttagcatgcagaataacgatggtggttttgcagcctttgaaaaaaatgtgggtaaacgcttttggcgctacctgctgatt
gaaggtgcagaatttctgctgatggacccgagtaccgcagatctgaccggtcgcaccctggaatatttcggcacctttgcaggtctgaccaaagatcatccggcagttgcccgtgcagttgatt
ggctgctggatcatcaagaggccgatggtagttggtatggtcgttggggtatttgttatgtttatggcacctgggcagcagttaccggtctgagcgcagttggtgttagtgccgatcatcctgcaat
gcaaaaagccgttcactggctgctgagcattcagaatgcagatggcggttggggtgaaagctgtaaaagtgatggtgcaaaaacctatgttccgctgggtgcaagcacaccggttcatacc
gcatgggcattagatgccctgattgcagccgcagaacgtccgacaccggaaatgaaagccggtgttcgtgccctggttcgtatgctgcaccatcctgattggaccgcaagctatccggttggt
caaggtatggcaggcgcattctatattcattatcatagctatcgctacatcttccctctgctggcgctggcacattatgaacagaaattttgtctgatcgccaattaa

10. *Mfu*SHC

atgcatagcggtcgtctgtttctgaaaaaagaaaatgaagtgggcgacaacaaaaaactgcatagcgttccgctgagcctggttgaagaaaccctgaattttccgcagaaagtggaaaaa
accatcaaaaaagcacagcgttatctgctgagcatccagaaagaagatggtcattgggttggtgaactgtttgttgatgttaccctggcatgtgattgtatccatctgatgcattggcgtggcaaa
atcgattacaaaaaacagaaacgcctggtgaaacatattctggatcgtcagctgccggatggtggttggaacatttatcctggtggtccgagcgaagttaatgcaaccgttaaagcatattttgc
cctgaaactggcaggttttagtccggatgaaccgctgatggcaaaagcacgtagcaccattctgcgtttaggtggtattccgaaatgtatgacctataccaaactgggtttagcactgctgggtg
tttatccgtgggatcgtctgccggttattccgcctgaaattattctgtttccgaactggtttccgttcaacctgtatgaaattagcgcatggtcacgcaccatgctggtgccgctgagcattattcatcat
tttaaaccgacacgtatcctgccggaaaaattacagctgcatgaactgttcccgtatggcaccgaacgtggtaaatttagctggctgaaaaagggtgcaaaatacctgagcaaagagggcc
tgtttctggcctgtgataaatttctgcagtattgggataaaaccagcctgaaaccgtttcgtaaaatggcaattgaaaaagccgagaaatggattctggaacgtattgcagcaggtagtgatggt
ctgggtgcaatctttccggcaatgcattatgcaattatggcactgattgcactgggttataccgaagataatccgattctgaagaaagccatcaccgattttgaaagcctggaagttgatgatcag
aaaaacgatgatctgcgtattcagccgtgtctgagtccgctgtgggataccgcaattggtctggttgcactggcagaaagcggttatgaacgtaatgcaccgcagcttaaaaaagcagcaca
ttggattattaaccgcgagattcgtattaaaggcgattggtatgttcgtaatccgcatccggaagcaagcggttgggcatttgaatataacaacatgtattatccggatgtggatgacacactgat
ggttctgctggcactgcgtctgattgatattgatgacaaaatcaaaaaagaagaggtgatgcagcgtgcactgcgttgggttattagctttcagtgtgaaaatggtggctgggcagcatttgata
aaaacgtgtacaaaaagtggctggaagatattccgtttgcagatcataatgcaattctggaccctccgtgtagcgatattaccgcacgtgcactggaactgttcggtaaaatgggtattcgcaa
aaacgaaaagtttgtgcagaaagcaatccgctatctgaaagaaacccaagaaagtgatggtagctggatgggtcgttggggtgtgaattatatctatggcacctggcaggccctgcgtggtc
tgcaggcaattggtgaagatatgaatcaagagtggattctgcgtgcacgtgattggctggaaagctgtcagaatgaggatggcggttggggtgaaacaccggcaagctatgataatccgca
gctgaaaggtaaaggtccgagcaccgcaagccagaccgcatgggcaattagcggtattatggcatgtggtgatatttttcgtccgagcattacccgtggtatcaaatatctgtgtgaacgtcag
ctgagtgatggttcatgggcagaagaatttctgaccggcaccggttttccgggtgtttttatctgaaatatgacatgtatcgcaatgcatggcctctgctggttattggtgaatattatcgtcagtatc
agcaggcaaaagaacgtgcaacctattgggttgatgcaaccctgggttgtatggaaaaacgtctgagcgcagtttaa

11. *Tel*SHC

atgccgaccagcctggcaaccgcaattgatccgaaacagctgcagcaggcaattcgtgcaagccaggattttctgtttagccagcagtatgccgaaggttattggtgggcagaactggaaa
gcaatgttaccatgaccgcagaagttattctgctgcataaaatctggggcaccgaacagcgtctgccgctggcaaaagcagaacagtatctgcgtaatcatcagcgtgatcatggtggttgg
gaactgttttatggtgatggtggtgatctgagcaccagcgttgaagcatatatgggtctgcgtctgctgggtgttccggaaaccgatccggcactggttaaagcacgtcagtttattctggcacgtg
gtggtattagcaaaacccgtatttttaccaaactgcatctggcactgattggttgttatgattggcgtggtattccgagcctgcctccgtggattatgctgctgccggaaggtagcccgtttaccattta
tgaaatgagcagctgggcacgtagcagcaccgttccgctgctgattgttatggatcgtaaaccggtgtatggtatggaccctccgattacactggatgaactgtatagcgaaggtcgtgcaaat
gttgtgtgggaactgcctcgtcaaggtgactggcgtgatgtttttattggtctggatcgtgtgttcaaactgttcgaaaccctgaacattcatccgctgcgtgaacagggtctgaaagcagcagaa
gaatgggttttagaacgtcaagaagcatcaggcgattggggtggcattattccggcaatgctgaatagcctgctggcactgcgtgcactggactatgcagttgatgatccgattgttcagcgtgg
tatggcagcagttgatcgttttgcaattgaaaccgaaaccgaatatcgtgttcagccgtgtgttagtccggtttgggataccgcactggtgatgcgtgcaatggttgatagcggtgttgcaccggat
catcctgcgctggtgaaagccggtgaatggctgctgagcaaacaaattctggattatggcgattggcacatcaaaaacaaaaaaggtcgtcctggtggctgggcatttgaatttgaaaatcgt
ttttatccggatgtggatgataccgcagttgttgttatggcactgcatgcggttaccctgccgaatgaaaatctgaaacgtcgtgccattgaacgtgcagttgcatggattgcaagtatgcagtgtc
gtccaggcggttgggcagcatttgatgttgataatgatcaggattggctgaacggtattccgtatggcgatctgaaagccatgatcgatccgaataccgcagatgtgaccgcacgtgttctgga
aatggttggtcgttgtcagctggcatttgatcgtgttgccctggatcgcgcactggcatatctgcgcaatgaacaagaaccggaaggttgttggtttggtcgctggggtgttaattatctgtatggca
ccagcggtgttctgaccgcactgagcctggttgcaccgcgttatgatcgttggcgtattcgtcgtgcagccgaatggctgatgcaatgtcagaatgcagatggtggatggggtgaaacctgttg
gagctatcatgatccgagtctgaaaggtaaaggtgatagcaccgcaagccagaccgcatgggcaattatcggtctgctggcagccggtgatgcaaccggtgattatgcaaccgaagcaat
tgaacgcggtattgcctatctgctggaaacccagcgtccggatggcacctggcatgaagattatttcaccggcaccggttttccgtgtcatttctatctgaaatatcactactatcagcagcattttc
cgctgacagcactgggtcgttatgcacgctggcgtaatctgctggccacctaa

12. *Afu*SHC

atgctgggtgcaattcgtgaaccgcctattgatgttcagattgcactgcatagccgtgatgataatcagacaggtctggttctgcgtggcacccgtcgtaccgttgatcgtgttctgaaaggtctgt
gtagcagcccgtgtttttttgtagcgttagcctgacaatggcaaccctgaccaccacaatggccaccaccgcaacgatggcaaccaccgaagcaagcaaaccgctggaagcacaggca
cgtaccgcactgaccaaagcaaccaattatgcatgggaaatctttagcaatcgtcattggtgtggtgaactggaaagcaatgttaccgttacctgtgaacacatcttctttctgtatgtgctgtacc
agcatattgatcctggtgaaggtagccagtatcgtcagtggctgctgagccagcagaatagtgatggtagctggggtattgcaccgaattatccgggtgatattagcaccagcgcagaagcat
atctggcactgcgtattattggtatgagcaccgatagtccggaactgtatcgtgcacgtacctttattcgtgcagccggtggtctgagcaaaatgcgtatgtttacccgtatctttttgccgaatttgg
tctggtgccgtggaccgcaattccgcagctgcctgcagaatttattctggttccggcacattttccgattagcatttatcgtctggcaagctgggcacgtagcaatgttgttccgctgctgattattgc
acatcatcgtccgctgtatccgctgccgaatggtctgcataaacagaatccgtttctggatgaactgtggttagatccggcaacaaaaccgctgccttatggtagcagcgatccgaccgatccg
gttgcatttgtttttaccattctggataaagccctgagctatttaggtggtctgcgtcgtagcccgacacgtggttatgcacgtcgtcgttgtgttcagtggattctgcagcatcaagaaaaagccggt
gattgggcaggtattattccgcctatgcatgccggtattaaagcactgctgctggaaggttataaactgcatgatgaaccgattcagctgggtttagcagcaattgaacgttttacctgggcagat
aatcgtggtaaacgtctgcagtgttgtattagtccggtttgggataccgttctgatgattcgtgcactgcaggatacaccggcaagcctgggtattaaacttgatccgcgtattgcagatgcactgg
catggaccgcagaaaatcagcatcgtggtccggaaggtgattggcgtgtgtataaaccgaacattccggttggtggttgggcatttgaatatcataatacatggtatccggacatcgatgatac
cgcagcagccgttctggcatttctgacccatgatcctgcaaccgcacgtagccgtctggttcgtgatgcagttctgtggattgttggtatgcagaatgcagatggtggctgggcagcatttgatcat
gaaaacaatcagctgttcctgaacaaaatcccgtttagcgatatggaaagcctgtgtgatccgagcacaccggatgttaccggtcgtaccattgaatgtctgggtatgctgcgtgatctgctgat
gcgtccggcagaaaatgccgaaaatggtgagaaatatggttatccggatggcgaaggtgatgcagcagcagacgcacatctgctgcagattattaacaccgcatgtgcacgtgccattcc
gtatctgattcgtagccaagaagcaaccggcacctggtatggtcgttgggcagttaattatgtttatggcacctgtctggtgctgtgtggtctgcagtatttcaaacatgatccgaaatttgcaccgg
aaattcaggcaatggcagcacgtgcagttaaatggctgaaacaggttcagaattccgatggcggttggggtgaaagcctgctgagctatcgcgaaccgtggcgtgcaggttgtggtccgtca
acaccgagccagaccgcatgggcactgatgggtattctgaccgtgtgtggtggtgaagatcgtagcgttcagcgtggtgttcgtcatctggttgatacccaggatgataccctgagccaaggt
gatggtggtgcagccgcatggacagaacgtgaatttaccatccgcgaaccgctgcatgaagcaagccagcgtattggtagcgattaa

13. *Apa*SHC

atgaatatggcaagccgtttcagcctgaaaaaaatcctgcgtagcggtagcgatacccagggcaccaatgttaataccctgattcagagcggcaccagcgatattgttcgtcagaaaccgg
caccgcaagaaccggcagatctgagcgcactgaaagcaatgggtaatagcctgacacataccctgagcagcgcatgtgaatggctgatgaaacagcagaaacctgatggtcattgggtt
ggtagcgtgggtagcaatgcaagcatggaagcagaatggtgtctggcactgtggtttctgggtttagaagatcatccgctgcgtccgcgtctgggtaaagcactgctggaaatgcagcgtccg
gatggtagctggggcacctattatggtgcaggtagcggtgatattaatgcaaccgttgaaagctatgcagcactgcgtagcctgggttatgcagaagatgatccggcagttagcaaagcagc
agcatggattattagcaaaggtggtctgaaaaatgtgcgtgtgtttacccgttattggctggcactgattggtgaatggccgtgggaaaaaaccccgaatctgcctccggaaattatctggtttcc
ggataattttgtgttcagcatctataactttgcacagtgggcacgtgcaaccatgatgccgctggcaattctgagtgcacgtcgtccgagccgtccgttacgtccgcaggatcgtctggatgccct
gtttcctggtggtcgtgcaaattttgattatgaactgccgaccaaagaaggtcgcgacgttattgcagatttttccgtctggcagataaaggtctgcattggctgcagagcagctttctgaaacgtg
caccgagccgtgaagcagcaatcaaatatgttctggaatggattatctggcatcaggatgcagatggtggttggggtggtattcagcctccgtgggtttatggtctgatggccctgcatggtgaa
ggttatcagtttcatcatccggttatggcaaaagcactggatgcactgaatgatcctggttggcgtcatgataaaggtgatgcaagctggattcaggcaaccaatagtccggtttgggataccat
gctgagcctgatggcattacatgatgcaaatgccgaagaacgttttacaccggaaatggataaagccctggactggctgctgagccgtcaggttcgtgtgaaaggtgattggagcgttaaact
gccgaataccgaacctggtggctgggcatttgaatatgccaatgatcgttatcctgataccgatgataccgcagttgccctgattgccattgcaagctgtcgtaatcgtccggaatggcaggca
aaaggtgttgaagaggcaattggtcgtggtgttcgttggctggttgcaatgcagagtagctgtggcggttggggagcatttgataaagataacaataaaagcatcctggcgaaaatcccgtttt
gcgattttggtgaagcgctggaccctccgagcgttgatgttaccgctcatgtgctggaagcatttggtctgctgggtctgcctcgtgatctgccgtgtattcagcgtggtctggcatatattcgtaaag
aacaggacccgaccggtccgtggtttggccgttggggtgttaattatctgtatggtacaggtgcagttctgcctgcactggcagcactgggtgaagatatgacccagccgtatattagtaaagc
ctgcgattggctgattaactgccagcaagaaaatggcggatggggtgaaagctgtgcaagctatatggaagttagcagcattggtcatggtgcaaccacaccgagccagaccgcatgggc
actgatgggtctgattgcagcaaatcgtcctcaggattatgaagcaattgcaaaaggttgccgctatctgattgatctgcaagaagaggacggtagctggaatgaagaagaattcaccggta

caggttttccaggttatggtgttggtcagacaattaaactggatgatccagccattagtaaacgtctgatgcagggtgcagaactgagccgtgcatttatgctgcgttatgatctgtatcgtcagctg
tttccgattattgcactgagtcgtgcaagccgtctgatcaaactgggtaattaa

14. *Bam*SHC2

atgattcgtcgcatgaataaaagcggtccgagtccgtggtcagcactggatgcagcaattgcacgtggtcgtgatgcactgatgcgtctgcagcagccggatggtagctggtgttttgaactgg
aaagtgatgcaaccattaccgcagaatatatcctgatgatgcacttcatggataaaatcgatgatgcccgtcaagaaaaaatggcacgttatctgcgtgcaattcagcgtctggatacacatg
gtggttgggatctgtatgttgatggtgatccggatgtgagctgtagcgttaaagcatattttgcactgaaagcagccggtgatagcgaacatgcaccgcacatggttcgtgcacgtgatgcaatt
ctggaattaggtggtgcagcacgtagcaatgtttttacccgtattctgctggcaacctttggtcaggttccgtggcgtgcaaccccgtttatgccgattgaatttgttctgtttccgaaatgggttccgat
cagcatgtataaagttgcatattgggcacgtaccacaatggttccgctgctggttctgtgtagcctgaaagcacgtgcccgtaatccgcgtaatattgcaattccggaactgtttgttacccctccg
gatcaagaacgtcagtattttcctccggcacgtggtatgcgtcgtgcatttctggcactggatcgtgttgttcgtcatgttgaaccgctgctgccgaaacgtctgcgtcagcgtgccattcgtcatgc
acaggcatggtgtgcagaacgtatgaatggtgaagatggtttaggtggtatttttccgcctatcgtttatagctatcagatgatggatgttctgggttatccggatgatcacccgctgcgtcgtgattg
tgaaaatgcactggaaaaactgctggtgacccgtccggatggttcaatgtattgtcagccgtgtctgagtccggtttgggataccgcatggtcaacaatggccctggaacaggcacgcggtgt
tgccgttccggaagcgggtgcaccggcaagtgccctggatgaactggatgcccgtattgcccgtgcctatgattggctggccgaacgtcaggttaatgatctgcgtggcgattggattgaaaa
cgcaccggcagatacccagcctggtggctgggcatttcagtatgcaaatccgtattatccggacattgatgatagcgcagttgttaccgcaatgctggatcgtcgtggtcgtacccatcgtaatg
cagatggtagccatccgtatgcagcccgtgttgcacgtgcgctggattggatgcgtggtctgcagagtcgtaatggtggttttgcagcatttgatgcagattgtgatcgtctgtatctgaatgccatt
ccgtttgcagatcatggtgcactgctggaccctccgaccgaagatgttagcggtcgtgttctgctgtgttttggtgttaccaaacgtgcagatgatcgtgcaagcctggcacgtgcgattgattatgt
taaacgtacccagcagcctgacggttcatggtggggtcgttggggcaccaattatctgtatggcacctggtcagttctggcaggtctggccctggcaggcgaagatccgagccagccgtatat
tgcgcgtgcactggcatggctgcgtgcccgtcagcatgccgatggcggttggggtgaaaccaatgatagttatattgatccggcactggcaggcaccaatgccggtgaaagcaccagcaat
tgtaccgcctgggctctgctggcccagatggcatttggtgatggtgaaagcgaaagcgtgcgtcgtggtattgcatatctgcagagcgttcagcaggatgatggtttttggtggcatcgtagccat
aatgcaccgggttttccgcgtattttctatctgaaatatcatggctataccgcctattttccgctgtgggcattagcccgttatcgtcgtttagccggtggtgttagcgcagcaggcgcacatgcagtt
ccggcaagtaccggtgcagatgcagcactggcctaa

15. *Mca*SHC

atgctgcgtgaagcaaccgcaattagcaatctggaaccgcctctgaccgcaagctatgttgaaagtccgctggatgcagcaattcgtcaggcaaaagatcgtctgctgagcctgcagcatct
ggaaggttattgggtttttgaactggaagcagattgtaccattccggcagaatatatcctgatgatgcactttatggatgaaattgatgcagcactgcaggccaaaattgcaaattatctgcgtag
ccatcagagcgcagatggtagctatccgctgtttcgtggtggtgccggtgatattagctgtaccgttaaagtttattacgccctgaaactggcaggcgatagcattgatgcaccgcacatgaaa
aaagcacgtgaatggattctggcacaaggtggtgcagcacgtagcaatgtttttacccgtattatgctggcaatgtttgagcagattccgtggcgtggtattccgtttattccggttgaaattatgct
gctgccgaaatggtttccgtttcatctggataaagttagctattggagccgtaccgttatggtgccgctgtttattctgtgtagccataaagttaccgcacgtaatccgagccgtattcatgttcgtga
actgtttaccgttgatccgcagaaagaacgccattattttgatcatgttaaaacaccgctgggcaaagcaattctggccctggaacgttttggtcgtatgctggaacctctgattccgaaagcagtt
cgtaaaaaagcaacccagaaagcctttgattggtttacagcacgtctgaatggtgttgatggtctgggtgcaatttttccggcaatggttaatgcctatgaagcactggattttctgggtgttcctcc
ggatgatgaacgtcgtcgtctggcacgtgaaagcattgatcgcctgctggtttttcagggtgatagcgtttattgtcagccgtgtgttagcccgatttgggataccgcactgaccagtctgaccctg
caagaagttgcacgtcatacagccgatctgcgtctggatgcggcactgagcaaaggtctgaaatggctggcaagcaaacaaatcgataaagatgcacctggtgattggcgtgttaatcgtg
caggtctggaaggcggtggttgggcatttcagtttggtaatgattattatccggatgtggatgatagcgcagttgttgcacatgcactgctgggtagcgaagatccgagctttgatgataatctgcg
tcgtgcagcaaattggattgcaggtatgcagagccgtaatggtggttttggtgcatttgatgcagataacacctattactatctgaacagcattccgtttgcagatcatggtgccctgctggaccct
ccgaccgcagatgttagcgcacgttgcgcaatgtttctggcacgctgggttaatcgtcagccggaactgcgtccggttctggaacgtaccattgattatttacgtcgtgaacaagaagcagacg
gtagctggtttggtcgttggggcaccaattatatctatggcacctggtcagttctgctggcgtatgaagcagccggtgttccgaatgatgatccgagcgttcgtcgtgccgttgcatggctgaaaa
gcattcagcgtgaagatggtggctggggtgaagataactttagctatcatgatccgtcatatcgtggtcgttttcataccagcaccgcgtttcagaccggttttgcactgattgccctgatggcagc
gggtgaagcaggtagtccggaagttcaggcaggcgttgattatctgctgcgtcagcagcgtccggatggtttttggaatgatgaatgttttaccgcaccgggttttccgcgtgtttttatctgaaat
atcacggctacgacaaatttttcccgctgtgggcattagcacgttatcgtaatgaacgttatgcactggcataa

16. *Sfu*SHC

atgcgtcgtctggatacctttccgcctgaaattccgaccggtagccgtgataaaccgcctagcggtgaagaacatagctgtagcacaccggcagaaccgctgcgtagccgtctggatgaag
gtattctgcgtgcagttgattggctggtttgtgatcagcatcctgatggtttttgggcaggtatgctgcagagcaatagctgtatggaagcagaatgggttttagccatgcattttctgggtattgatga
tgatccgaaatatgatggtgtgattcgtgcaattctgggtgaacagcgtgcagatggtagctggggtgtttttcataaagcaccgaatggtgatatcaataccaccgttgaatgttatgcagcact
gcgtgcaagcggtctggcaccggaaagcgcaccgctgagcagcgcacgtgaatggattctggcaggcggtggtctggcaaatattcgtaattttaccaaatattggctggccctgattggtg
aatggccgtgggaaggcaccccgaccattcctccggaactgatctttttccgcctcgtatgccgctgaacatttatcattttgcaagctgggcacgtagcaccattgttccgctgagtattctgag
cgcacgtcgtccggttcgtccgctgccggaagatcgtcgcctggatgaactgtttccgcagggtcgtagcgcatttgattttcgtctgcctcgtaaagatggttggctgagctgggaaggtttttttc
atgtttgcgatcgtatcctgcgtctgtatgcacgtacccgtcgtgcaccgtttcgtgaaaccgcaattcgtgtttgtctggaatggattattcgtcgtcaagaaaccgatggtgcatggtcaggtattc
agcctccgtggatttatgcactgctggcactgcatgccgaaggttatggtctggatcatccgatcctgcgtgccggtctgcgtgcctttgatagccattggagctatgaacgtgatggtggtatttat
ctgcaggcaagcgaaagtccggtttgggataccgttctgagcctgcgtgcactggcagattgtggcgaagaacgtaaagcaagcgttagcattgcaagcgcactggaatggctgctgaatc
gtcagattagcgttcctggtgattgggcagttcgtgttccgagcgttccgtgtggtggttgggcatttcagcgtgcaaatagcttttatccggatgttgatgataccgcagttgcaattgaagttctggc
acgtctgcgtccgtttaccgcaaatcagagcgcagttgatcgtgccattcgtagtgcacgtgattgggtgttagcaatgcagtgtagcaatggtggctgggcagcatttgatcgtgataatgatttt
aaactggtgaccaagattccgttttgcgattttggtgaactgctggacccaccgagcgttgatgtgaccgcacatgttattgaagccctggcagcattaggttgggatatgaccagccgtgaaat
tgaagcagcagttagctttattcgccgtgaacaagaagccgaaggtagctggtttggtcgttggggtgttaatcatatttatggcaccgcaaccgttctgcctgcgctgcgtgccattggtgaaga
tatgagcagtgcctatgtgctgcgtgcggcagactggctggcaagccgtcagaatgcagatggcggttggggtgaaacaccggcaagctatatggatgatagtctgcgtggtgttggtgaaa

gcaccgcatcacagaccgcatgggcaattatgggtttagttgcagttggtagcggtgcacatgatgatacagttcgtcgcggtattgattttctgctgtttgcacagcatggtggcacctgggaagaaccgcagtataccggcaccggttttccgggttatagcgttggtgaacgtattcgtctgcgtgatatgggtgcaagcctgaaacagggcaccgaactgcagcgtgcatttatgattaattataacctgtaccgccactactttccgctgatggcactgggtcgtgcacgttatcatctgcagctgcgtcgttcagcacgcgaaggtggtaatggtgaaacaaccccgaatggtagcgcactgtaa

17. *Ttu*SHC

atggaaatccaggatgaagttgatctgctggaaccgcaagaaagcctgaccgcaagcgcagatagcgcagttgatcgtgcactgttttggctgctggatgcacagtatgaagatggttattgggcaggtattctggaaagcaatgcatgtatggaagcagaatggctgctgtgttttcatgttctgggtattgcaaatcatccgatgagccgtggtctggttcagggtctgctgcagcgtcagcgtgcagatggtagctgggatgtttattatggtgcacgtgccggtgatattaacaccaccgttgaagtttatgcagcactgcgttgtcagggttatgcagccgatcatccggatattaaacgtgcccgtgattggattcagctgcaaggtggtgtgaaacaggttcgtgtttttacccgtttttggctggcactgattggtgaatggccgtgggaagaaaccccgaatctgcctccggaaattctgtttttccgcgttggtttccgttcaacatttatcattttgcagcatgggcacgtgcgaccctggttccgctgtgtattctgagcgcacgtcgtatggttgttccgctgaacaaaaaaagctgtctgcaagaactgtttccggaagatcgtagcgcagtggttgcactgggtaaaaaagccggtgcctggtcaaccttttctatcatgcagatcgtgccctgaaaaaataccagcgtacctttaaacgtccgcctggtcgtcagcaggccattaaaatgtgtctggaatggattctgcgtcgtcaggatgccgatggtgcatggggtggtattcagcctccgtggatttatagcctgatggcactgaaagcagaaggttatccggttacacatccggttatggcaaaaggtctggcagcactggatgcccattggagctatgaacgtcctggtggtgcccgttttgtgcaggcatgtgaaagtccggtttgggataccctgctgagcagctttgcactgctggattgtggttttagctgtaccagcagcagcgaactgcgtaaagcagttgactggattctggatcagcaggttctgctgcctggtgattggcagcaaaaactgccgaccgtttcacctggtggttgggcatttgaacgtgccaatgttcattatcctgatgttgatgataccgcagttgccctgattgttctggcaaaagttcgtcctgattatccagataccgcacgtgttaatctggccattgaacgtggtctgaattggctgtttgcaatgcagtgtcgtaatggtggctggggtgcatttgataaagataacgataaagacctgctgaccaaaattccgtttagcgattttggtgaaaccattgatccggcaagcgttgatgtgaccgctcatgtgctggaagcactgggcctgctgggttatcgtacaacccatccggcagttgcaaaagcactggaatttattcgtagcgaacaagaaaatgatggttgctggtttggtcgttggggtgtgaattatatctatggcaccgcagcagttttaccggcactggcaagcctgaatatgaacatgaatcaagaatttatccgtcgcgcagcaaattggattttaggcaaacagaataatgatggcggttggggtgaaagctgtgcaagctatatggatgatacccagcgtggtcgtggtccgagcaccgccagtcagaccgcatgggcaatgatgagcctgctggcagttgatggtggcacctatgccgaaagcctgctgcgtgcagaagcatatctgaaaaccacacagacaccggaaggcacctgggatgaaccgtattacaccggcaccggttttccaggttatggtattggtcgtcgtgaaattaaacgtcagcgtagcctgcagcagcatgcagagctgagtcgtggttttatgattaattataacctgtaccgccactactttccgctgatggccctgggtcgcctggcagccctgcgtggtgcataa

18. *Aca*ACH

atgctgccgtataatcaggatcatcattttggtaaagttgccgaaaatgcaaccatgcctccgacactggatgaagcaattgaacgtagccaggattttctgctgagcctgcagtatccggaaggttattggtgggcagaactggaagcaaatgttaccctgaccgcacagaccattatgctgtacaaaattctgggcatcgatcataaatacccgatccacaaaatgaaaacctatattctgcgtacccagcgtgcacatggtggttgggaaatcttttatggtgatggtggctgtctgagcaccaccattggtgcatatatggcactgcgtattctgggtgttccgaaaaccgatccggtgctgcagaaagcactgaaactgattcatagcaaaggtggtgttaccaaaagccgcatgtttacaaaaatttgtctggcactgctgggttgctatgattggaaaggtattccgagcctgcctccgtggctggttctgctgccgagctggtttccgtttagcctgtatgataccgcaagctgggttcgtggttgtgttgttccgctgaccattatctttgataaaaaaccggtgtacaaactgaatccgctgctgtgtctggatgaactgtatagcgaaggtaaaggtaaagcacgtgttcacctgagctttattccaggtgattggaccagcaactttttttgttggtctggatcacgtgtttaaatacatggaaaatctgggcgttgttccgtttcgtcagtggggtattaaagaagcagaacgttggaccctggaacgtcatgaagatagcggtgattttcatggcatttatccgcctatgttttatagcattgttagctatagcctgctgggctatgaaattacagatccggttgttcatcgtgcactggaaagcatgcgtggttttaccgttgaacgtgaagatgaatgcgttgttcagagctgtattagcccgatgtgggataccgcatttgttattcgtagcctggcagaaagcggtctgcagccggatcatcctgcactgcaaaaagccggtgaatggctgctgcaaaaacaggcgacccagcatggtaattggttttataagaaacgtaccggtcgtgcaggcggttgggcatttcagtttttaaccgttggtatccggatgttgatgatagcgcagcagttagcatggcactgaatgcaattaaactgcaggacgatgatgttaaaaaaggtgccattaaacgttgtgccgaatggattagcgttatgcagtgtaaagatggtggatgggcagcctatgattgtgataatgatcgcgaatggctgaattgtaccccgtttggtgatctgaaagcaatgattgatccgaataccgttgatgttaccgcacgtgtgctggaaatggttggtcgtgtgaaagaggcaggcgacgcaagcgcaattctgcctccgcgtgcaattgcccgtggtctggcatatctgcgtcgtgaacaagaaaccgaaggttgttggtatggtcgttggggtgtgaattatatctatggcaccagcggtgcactgatggcactggcactggttgcaccgagcacacataaagaagaaatcgaacgcggagcacgttggctggttgaagttcagaataaacgtggcaccaaaggtgcaaatggttatagccataccaatggcgcacgtgaaggtggcgttgcaatgaatggcaattgtaaaaacatgggtgcaccggaagatggcggttggggtgaaacctgttttagctataatgatattaccctgaaaggtcgcaatgaagtttcaaccgttagccagaccgcatgggcactgcagggcctgctggcagccggtgatgcactgggtaaatatgaagttgaaagcattgaacatggcgtgcagtatctgctgtcaacccagcgcaaagatggtagctggtgtgaaaaacattttaccggtggtggttttccgcgtttttctatattcgttatcatctgtatgccggtcattttccgctgagtgccctggcacgttatcgtgatcgtgttcgtgcaggtaaaatggccaaataa

19. *Bsu*TC

atgggcaccctgcaagaaaaagttcgtcgttttcagaaaaaaaccattaccgaactgcgtgatcgtcagaatgcagatggtagctggaccttttgttttgaaggtccgattatgaccaacagcttttttatcctgctgctgaccagcctggatgaaggtgaaaatgaaaaagaactgattagcagcctggcagcaggtattcatgcaaaacagcagccggatggcacctttattaactatccggatgaaacccgtggtaatctgaccgcaaccgttcagggttatgtgggtatgctggcaagcggttgttttcatcgtaccgaaccgcacatgaaaaaagcagaacagtttattatcagtcatggtggtctgcgtcatgtgcattttatgacaaaatggatgctggcagcaaatggtctgtatccgtggcctgcactgtatctgccgctgagcctgatggcactgcctccgacactgccgattcatttttatcagtttagcagctatgcccgtattcattttgcaccgatggcagttaccctgaatcagcgttttgttctgattaatcgcaatatcagcagcctgcatcatctggaccctcacatgaccaaaaatccgtttacctggctgcgtagtgatgcatttgaagaacgtgatctgaccagcattctgctgcattggaaacgtgtttttcatgcaccgtttgcatttcagcagctgggcctgcagaccgcaaaaacctatatgctggatcgtattgaaaaagatggcaccctgtatagctatgcaagcgcaaccatttatatggtttatagcctgctgagtctgggtgttagccgttatagcccgattattcgtcgtgcaattaccggtattaaagcctggttaccaaatgcaatggtatcccgtatctggaaaatagcaccagcaccgtgtgggataccgcactgattagttatgcactgcagaaaaatggtgttaccgaaaccgatggtagcgttaccaaagcagccgattttctgctggaacgtcagcataccaaaattgcagattggagcgttaaaaatccgaatagcgttcctggtggttggggttttagcaatatcaataccaataatccggactgtgatgataccaccgcagttctgaaagcaattccgcgtaatcatagtccggcagcatgggaacgtggtgttagctggctgctgagcatgcagaataatgatggtggttttagcgcctttgagaaaaatgttaatcatccgctgattcgtctgctgccgctggaaagcgcagaagatgcagcagttgatccgagcacagcagatctgaccggtcgtgttctgcattttctgggtgaaaaagttggctttaccgaaaaacatcagcatattcagcgtgcagttaaatggctgtttgaacatcaagaacagaatggcagctggtatggtcgttggggtgtttgttatatctatggcacctgggcagcactgaccggtatgca

tgcctgtggtgttgatcgtaaacatccgggtattcagaaagcactgcgttggctgaaaagcattcagaacgatgatggttcatggggtgaaagctgtaaaagtgcagaaatcaaaacctatgtt
ccgctgcatcgtggcaccattgttcagaccgcatgggcattagatgcactgctgacctatgaaaacagcgaacatccgagcgttgttaaaggtatgcagtatctgacagatagcagcagcca
tagcgcagatagcctggcatatccggcaggtattggtctgccgaaacagttttatatccgctatcatagctatccgtatgtttttagtctgctggccgttggtaaatatctggatagcatcgaaaaag
aaaccgccaacgaaacctaa

20. *Gni*PNG

atggcactgccgtttaatcaggatagctataaaggtgatgatgaagccgatgttagcaaaggtgcagcaaaaagccctccgagcctggaagaagcaattcagcgtagccaagaatttctgc
tggcacagcagtttccggaaggtttttggtttggtgaactggaagcaaacgttaccattattagccataccgtgatcctgtataaactgctgggtatcgaagaaaacttcccgatgtataaattcga
acgttatctgcgtcgtatgcagtgtagtcatggtggttgggaaattgcctatggtattggtagctatctgagcgcaaccattgaagcatatattgcactgcgtctgctgaatgttccgcagagcgatc
cggcactgcagaaagcactgcgtgttattctggatagcggtggtgttaccaaagcacgtatttttaccaaaatttgtctggccctgctgggttcatttgattggcgtggtattccgagtctgcctccgt
ggctgattctgtgtccgacctggtttccgctgagcatttatgaagttagcagctgggcacgtggttgtattgttccgctgctggttatcctggataaaaaaccggtgtttaaagttagtccggaagtga
gctttgatgaactgtatgccgaaggtcgtgaacatgcctgtaaaatcattccgattagcggtgattggaccagcaaatttttcattaccgttgatcgcgtgttcaagatgatggaacgtctgcgtgtt
gttccgtttcgtcagtggggtattcgtgaagcagaaaaatggattctggaacgtcaagaggaatcaggtgattacgttaacatttttccggcaatgttctatagcgtgatgtgcatgaaagttctgg
gttatgaaaccaccgatccggttgttcagcgtgcactgctgggctttaaaggttttaccattgaaaccgcagatgagtgtaaagttcagagcaccgttagtccgatttgggataccgcatttattgtt
cgtgcactggttgatagcggtattccgcctgatcatcctgcgctgcaaaaagcaggtcagtggctgctgcagaaacaaattctgaaacatggtgattgggcctttaaagatcgtcagaatccg
gtgaatcagcgtggttttgcatgtctgcagcgtgatagccagattgaaacagccgatgaatgtcgtgtgcagagcaccctgtcaccggtgtgggatacagcctttgttgttaaagccctggttgat
tcaggtattcctccgaaccatccggctttacagaaagctggccagtggttactgcaaaatcagaccctgacgcacggcgattgggcattcaaaacccagagcggtcatctggcagcaggcg
gttgggcgtttcagagccataatcgttggtatccggatgcagatgatagcgcagcagttatgatggcactggattgcattgaactgccggatgaagatgttaaaaatggtgcaattgcccgtggt
ctgaaatggattagcgcactgcagtcaagaaatggtggctgggcaggttatgataaaaactgtgatcagcagtggattaacaaagtgccgttcaatgatctgaatggcatcctggatgttccg
acagcagatgttaccgcacgtgttctggaaatggttggtcgtctgagccgtctgggtgcagttggcaccccgtatagtccgcgtcattgtaccctggtggaaagcattccgcatctgctgctgccg
gaaaccattgcacgcggtctggcatacctgcgtcgcgaacaagagggtgaaggttgttggtggggtaaatggggtgtgaattatatctatggcacctgtggtgcgctgctggccctgagccag
gttgcaccgaccacacatcaagaagaaatcgcacgcggagcaaaatggctggcccaggttcagaatcgttgtgataaacagaaagcagcacagggtccgcgtgatggtggatggggtg
aaagctgttttagctatgatgatcctgcactgaaaggtcagaatgatgcaagcaccgcaagtcagaccgcatgggcagtgcagggcctgctggcagccggtgatgcactgggtaaatatga
agttgaagcaattgaacagggtgtgcagtatctgctggcgacccagcgtaaagatggtacatggcatgaagcacattttaccggtagctgttttgcccagcatttttatgtgcgttatcactattatg
cgcagcattttccgctgtcagcactgggtctgtatcgtacccgtatcctgcagcatcagtaa

21. *Bte*SHC

atgaacagcgaactggaacgtctgaccgcagttctgcagcgtgaacagcaggcagatggtagctggcgttattgttttgaaagcggtccgctgaccgatgcctatgcaattattctgctgcgta
ccctgaatattccgaatgaaccgctgattagcggtctggcaaaacgtattgcaagccgtcaggcaccggatggcacctggaaactgtatagtgatgaacgtgatggtaatctgagcaccacc
attgaaagctattttgcactgctggcagccggtgcagcagcaccgtcagatgaacgtctgcaggcagcacgtcgttttattcgtgcaaaaggtggtctggcacaggcaaatctgggcacccgt
gttatgctggcactgacaggtcagcatccgtggccaccgtttccgattccggcagaatttatgctggttccgcctttttttccgctgcacctgtttgatctggttggttttgcacgtgttcatctggttccga
ttatggttgcagccgcacgtacctttgcagttcgtagtcgtcagatgccggatctgtcagacctgtttcgtggtctgccgagcgcaggtccgcagcagctggatctgatttggtttcatgaactgatt
gatagcggtattcgtagcctgccgagctttctgcgtccggcacgtgaactgggtctgcgtgaagcagaacgttttctgctggatcgtctggaaccggatggtacactgtatagctttttttaccgcaa
cctttctgatgattttcgccctgctggccctgggttatcgtccggatcatccggttattattcgcgcagttcgtggtgcagaacgtctggtttgtccggttggtgatgttctgcacatgcagaatagcac
cagcaccatttgggataccgcactgctgagccatgcactgcagaccgcaggtatgccggttagtcatccggtgattcagcaggcaacccgttatctgctgagtcgtcagcataatcgttatggt
gattgggcacgtcgtagtccgggtgttccgcctggtggttggggttttagcgatattaacaccattaatccggatgtggatgataccaccgcagcactgcgtgcactgcatcgtgcaggtagcgg
tgatccggcaattcgtcaggcatgggatcgcggtctgcgttggctgctgagtatgcagaattcagatggcggttggcctgcatttgaacgtaataccgcaaatccgctggttaaactgttaccgg
caggcggtgccgaagcagcatttaccgatccgagcaccgcagatctgaccggtcgtaccctggaatttctgggtaatcatgcaggtatgaccctgcagcatccggcagtgcgtcgtggtgttg
actggctgctgcgtcatcaagaaaccaatggtagttggcatggtcgttggggtatttcatatctgtatggtacatgggcagcactgagcggtctgattgcagcgggtgttagccctgatcatcctg
caattcagaaaggcgttagctggctgcgtagcgttcagaatcgtgatggtggctggggtgaaagctgtcagtcagatgttctgaaacgttatgttccgctgggtgcaagcaccccgagtcaga
ccgcatgggcagttgatgccctgacagcagttagccgtcgtaccggtccggaactggaagccggtgttcgttttctgttagcagcaggtaaacgtcgtgattggaccagcagctatccgaccg
gtgccgcactgccaggtggcttttatatccattatcatagctatcgctatatctggcctctgctgaccctggcacagtatcgtaacaaatttcagccgtaa

22. *Cth*SHC

atgcctggttttgcaccgcgttttgttcagccggttgttgaaagtccgctgcctccggcatttcgtagcgcacgtccggcaccggcaaccgcagcagcagttgaagcagcaattcgtaaagcac
aggcatatctgctgagcaaacagtatccggaaggttattggtgggcagaactggaagcaaatgttaccctgaccgcagaatatgtgtttctgcataaagttctgggcaccgatggtgaacgta
cccgtcagtttgaaaaaattcgtacctatctgcgtcgtcagcagcgtgaacatggtggttgggaactgtattatggtgatggtggtgaactgagcaccagcattgaagcctattttgcactgaaac
tgctgggtgatagtccggatctgccgcacatggcacgtgcgcgtcagtttattctggcacgtggtggtattaccaaagcacgtgtttttaccaaaattcacctggcactgtttggtgcatttccgtgg
gaaggttgtccgacactgcctccgtggattatgctgctgccggattggtttccgtttaccatttatgaactggcaagctgggcacgtagcagcaccgttccgctgctgctggttagcgatcgtaaa
ccggttgttcgtgttcctggtggtgatgcagatgaactgtatgccgaaggtcgtgcacaggccgatctgagcctgccgaatcctgcaggtctgctgagtttaggtggtgtttttattggttttgactgg
atgctgaaactgatggaacgttttgatctgagtccgcgtcgtgccgaagcactggcacgcgcagaacagtggaccctggaacatcaggatgatagcggtgattggggtggcattattccggc
aatgctgaatagcctgctgggtctgcattgtcgtggttatgcaccgacacatccggtaatgcagaaaggtattgcagccgttgaacgtttttgtatcgaaaccgaagatgaatttcatacccagc
cgtgtgttagtccggtttgggataccggtctgaccattctggccctgctggatagcggtctgccgaacgatcatccggcactggttcgtgccggtgaatggctgctgtcaaaacaaatttttcgtga
tggtgattggcgctttaaaaaccgtaccggtccggcaggcggttgggcatttgaattctggaatgatttttttccggatgtggatgataccgcagttgttacaatggcactgcatcgtttaaaactgc
ctgatgaagcagaaaaacagcgtcgtctgaaactggcaattgaatggaccctgagcatgcagagcaaaaatggtggctggggtgcatttgatgttgataataccctggaaatcctgaacga

tattccgtatggtgatctgaaagcaatgattgatccgcctaccgcagatctgacaggtcatattctggaaatgctgggcgttaccggttatgcagcaccgcgtgaaaaagttgaacgtgcaattg
cctttatcaagagcaaacaagaacctgaaggctgttggtggggtcgttggggtgtgaattacatttatggcacccacatggttatttgtggtctggttgcactgggtttagatccgcgtgaagccttt
attatgcgtggcacccagtggctgaatagttgtcagaatgaagatggcggatggggtgaaacctgtgcaagctatggcgatcgtaccctgatgggtgttggtaaaagcaccccgagccaga
ccgcatgggcactgctgggactgatggctggcggtgaaggtaaaagcgattgtgcccgtcgtggtattgaatatctggttacccatcagaacgatgatggtagctggaccgaagcagaattta
ccggtacaggctttccgaatcacttctatatgaactatcacttttaccgcaactactttccgctgatggcactgggtcgttatcgtgcatttgcacgtacctaa

23. *Ctt*SHC

atgaatctgagcggtcaggttgatcaggcagttggtcgtctgagcgaaagcctgagccgtatgcagagtgatgatggtagctggcgtttttgttatgaaaatgcagttctgaccgatgcctatatg
attattgcactgcgtaccctggaaattgcagatgaaccgctgattcgtcagctgcgtgatcgtctgctggcaacccagtatgcagatggtgcatggcgtgcatatccggatgaacgtgaaggta
atctgagtgcaaccgttgaatgttattatgcactgctgtatagcggttatagccgtgatgccgatccgcctctggttaaagcccgtgcatttattctggcaaaaggtggtattcagcagattggtggt
ctgctgaccaaagttatgctggcaagcaccggtcagtatccgtggcctcgtagcctgaaaatcccgctggaatttctgctgctgccgagccagtttccgctgagcgtgtttgattttagtggttatg
cacgtgttcacatgattccggcactgctgctggcagatcgtcgttttagcctgcgtaccaaaaccagtccggatctgagtgaactggcaggcgatcgtagccgtgaaccgcctagctggtttgat
ccggcattaggtcgtggtcatccgcgtgaactgcagagcctgctggaacaaattggtaaaggtattgaacgcctgaatggtattccgagccagctgcatgaagaagcagttcgtcgcgcag
aacgttatatgctggaacgtattgaagcagatggcaccctgtatagctatgcaaccagcacctttctgatgatcctggcgctgctggccattggttatgataaacgtcatgcagttattcgtaatgc
cgttcagggtctgcgtgcaatggtttgtcgtggtgcacagccgatttttctgcaaaatgcaccgagcaccgtttgggataccgcactgctgagcagcgcactgcaagaggcaggcgcagatg
caaatagcccgatgattcgtcgtgcaaatgcatatctgctggcgaaacagcatcgtaaacctggtgattggctggttcataatccgagcgcagttccaggtggttggggttttagcgataccaat
accattaatccggatgttgatgataccaccgcagcactgcgtgccattaaacgtcaagccggtgcagatcctgcatatcgtgaagcatggaatcgtggtctgcattggctgctgtctatgcaga
acgatgacggtggttggcctgcatttgaaaaaaacaccgatcgtcagattctggttctgctgccgctgcgtgaagcaaaaagcagcgcaattgatccgagcaccagcgatctgaccggtcg
caccctggaatttttaggtaattatgcaggtttaggtatgggccatgcctttattcgccgtggcaccgattggctgattggtcatcaagaaaaagatggttcatggtatggtcgttggggtgtttgttat
atctatggcacctgggcagcactgacaggtctggcagcagccggtattcgtgcagatcatccggcagttcgtgccggtgcacagtggctgaccgatattcaacaggccgatggtggctggg
gtgaaagctgtgatagcgatcgccagatgcgttatattccgctggaagaaagcaccccgagccagaccgcatgggcattagatgcactgattgcagttcatgaagcaccgacaccgacca
ttgatcgtggtattcgtcgtctgatcggcctgctgcaagaagaaagccgttttgcagcatatccgaccggtgcaggtctgcctggtatcttttatagccactattatagctatcgctacatctggcctc
tgtttgcactggcacactataaaagcaaatatagcagctaa

24. *Gth*SHC

atggcaggcgaacgtagcgcactgattaccgcactgaaacgtagccaggcagcagatggtagctggcgttttccgtttgaaaccggtattagcaccgatgcctatatgattattctgctgcgta
ccctggatattaatgatgaaccgctgattcaggcactggttgaacgtattgaaagccgtcaagaggcaaatggtgcatggaaactgtttgcagatgaaggtgatggtaatgttaccgcaaccg
ttgaagcatactatgcactgctgtatagcggttatcgtcagccgaccgatcgtcacatgcagaaagcaaaacgtcgtattctggatatgggtggtctggatcgtattcacctgtttaccaaagttat
gctggcactgacaggtcagtatccgtggccaggtcgttttccgctgccgctggaatttttctgctgccaccgagctttccgctgaatatgtatgatctgagcgtttatggtcgtgccaatatgattcc
gctgctgattgcagcagatagccgttatagccgtaaaaccgataaaagtccggatctgagtgacctgtttgcaagccgtggtgattggggtatgccggaaagccgtagcctgctgacctatgtt
aaacgtagtctgattggtctgcctgcacagctgcatcaggcagcaaaacagcgtgcagttcgttacctgtttgaacatatcgaaccggatggcaccctgtatagctattttagcagcacctttctg
tttatctttgccctgctggccctgggttatcgtaacgatgatccgcgtattcgtcaggcagttcgtggtctgcgttcactgcgtaccaccattgatggtcatgtgcatctgcagtataccaccgcaagc
gtttggaatacagcactggcaagctataccctgcaagaggcaggcgttccgatgaccgatcgcgcaattgaaaaagcaaatcgttatctgctgagccgtcagcatgttcgttatggcgattgg
gcagttcataatccgtatagcacccctggtggttggggttttagtgatgttaataccatgaatccggatgtggatgataccacagcagcactgcgtgcaattcgtcaagcagcagcaaaagaa
accgcatttcgtcatgcatgggatcgtgcaaatcagtggctgtttagcatgcagaatgatgatggtggttttgcagcctttgaaaaaaatgttagcagccgtttttggcgctatctgccgattgaag
gtgcagaatttctgctgatggacccgagcaccgcagatctgaccggtcgcaccctggaatatttcggcacctttgcaggtctgaccaaagatcagcgtgccgttagccgtgcagttgattggct
gctgagtcatcaagaacgtaatggtagttggtatggtcgttggggtatttgttatatctatggcacctgggcagcaattaccggtctgaccgcagttggtgttccggcacatcatccggcactgca
aaaagccgttcgctggctgctgtcaattcagaacgatgacggtggctggggtgaaagctgtaaaagtgatggtgcaaaaacctatgttccgctgggtgatagcacaccggttcataccgcat
gggcattagatgccctggttgcagcagcagaacgtccgacaccggaaatgaaagcaggttttcgtgcactgtttcgtctgctgcatcatcctgattggaccgcaagctatccggttggtcaagg
tatggctggtgccttttatatccattatcatagctatcgctatatcttcccgctgctggcgctggcacattatgaacagaaatttggtccgctggatgactaa

25. *Sth*SHC

atggaccctgcactgagccgtgcagttgattggctgctggaacatcaagatccggcaggttggtggtgtggtgaatttgaaaccaatgttaccattaccgcagaacatattctgctgctgcgtttt
ctgggtttagatccgagtccgctgcgtgatgcagttacccgttatctgctgggtcagcagcgtgaagatggtagctgggcactgtattatgaaggtccggcagatctgagcaccagcattgaag
cctatgcagcactgaaagttctgggccttgatccgaccagcgaaccgatgcgtcgtgcactgcaggttattcatgatttaggtggtgttgcacaggcacgtgtttttacccgtatttggctggcaat
gtttggtcagtatccgtgggatggtgttccgagcatgcctccggaactgatttggctgcctccgagcgcaccgtttaatctgtatgattttgcatgttgggcacgtgcaaccattacaccgctgctga
ttattctggcacgtcgtccggttcgtccgctgggttgtgatctgggtgaactggttctgcctggtagcgaacatctgctgacacgtgttccaggtagcggtccgttttggtggggtgataaagttctga
aacgttatgatcatctggttcgtcatccgggtcgtgatcgtgcatgtcagcgtattgttgaatggattattgcacgtcaagaagcagacggtagttggggtggtattcagagcgcatgggttatgag
cctgattgcactgcatctggaaggtctgccgctggatcatccggttatgcgtgcaggtctggcaggttttgatcgtgttgcactggaagatgaacgtggttggcgtctgcaggcaagcaccagtc
cggtttgggataccgcatgggcagttctggcactgcgtcgcgcaggtctgcctcgtgaacatccgcgtctggccctggccgtggactggctgctgcaagagcagattcctggtggtggtgattg
gcaggttcgtaccggcaccattccaggcggtggttgggcatttgaatttgataacgatcattatccggacatcgatgataccgcagttgttgttctggctctgctggaagcaggtcatgaagatcg
tgttcgtaatgcagttgaacgtgcagcccgttggattctggccatgcgtagcaccgatggtggctggggtgcatttgatcgcgataatgcccgtgaagttattcatcgtctgccgattgcagattttg
gcaccctgattgatccgcctagtgaagatgttaccgcacatgttctggaaatgctggcacgcctgagctttccgagcaccgatccggttgttgcccgtggtctggaatttctgcagcagacccag
cgtccggatggtgcatggtttggtcgttggggtgtgaattatatctatggcacctggtgtgcagtttcagcactgaccgcatttgcagataccgatgcaaccgcacgcgcaatggttccgcgtgcc
gttgcctggctgttagatcgtcagaatgcagatggcggttggggtgaaacctgtggtagctatgaagatccgaatctggcaggcgttggtcgtagcaccccgagccagacagcctgggcagt

tttagccctgcaggcagccggtctgggccagcatcctgcatgtcgtcgcggtctggatttcctgcgtgaacgtcaggttggcggtacatgggaagaacgtgaacacaccggcaccggttttcc
gggtgatttttcattaactatcatctgtatcgccacgtgtttccgacaatggcattagccggtgcagcaaccggtatggatagtccgcgttaa

26. *Tsg*SHC

atgcagagccagtggattctgcatgttcaggcaatgattcatcgtctgaaaaaagagctgatccagaaacagaatccggatggcacctggccgttttgttttgaaaatggtattggcaccgatg
cctattatgttctgctgcatcaggttctgaaacgtcctgatccgggtacactggcaccggttctggaacgtattctggataaacagaccggtgatggtacatggaaagcatttccggatgaaaaa
gaaggtaacgttagcgcaaccctggatgcaagcctggcactgctgtatagcggtgttaaaacaccggatgatccgagcctgaaacgtgcccgtgattttctgctgagccgtggtatggaaac
caaagcaggtagtctgacccaggttgttctggccctgctgggtcatcgtagctggtcacgtattaccaaactgccggttgaatttttctgcttccggcagcaagtccggttaactttttgattttgttg
gttatgcccgtgtgcatattgcaccggttatgctggcaagcgatcaggatttctatattcatctgaaaggttatcgcgaggttgaagattggctgccgagcagctttcgtacctatatggaacgcat
gcatccggactatttttggaccaaagaagatctgctgcctgcagttgaaaccaccgcatatagcacctttttaagaaaaatgttcatcagcgtgccctgtattggggtgaaaactttctgctgtctc
gtattgaagaagatggcaccctgtatagttatatgaccagcacctttctgatgatttttgcactgctgagtctggattatccgcctgatcatccgctgattcagaaagcaatggaaggtctggatcgt
atgatttttccgctgcaagaaggtgcacatctgcaagaggcaaccagcaccgtttgggataccagcctggttatgaccgcactgcagaatgcaggtctgagccctggtcatggtgttattcaaa
aaggtcgtaattacctgctgctgaaacagcataccagctgtggtgattggtgcctgaaaaatcgttatgcaattcctggtggttggggttttagcgataccaatacaattaacccggatgttgatg
ataccgttcagtgtctgcatgcaattgctccggcagttcgtgaaggttgggcacaagatgaatggaaacgtggtctgcagtggctgctgagtatgcagaaccgtgatggtggctggcctgcattt
gaacgtaataccaataaaatgtggctgaaactgctgccagcacgtaatgaaaaacgtgtttggggtgatccgagtaccgcagatctgaccggtcgtgttctgcattttctgggtagcgaattag
gttggaccattgatcgtccggaagttcgtcgtgcatggtcatggctgtatcatcaccagaattcagatggtagctggtttggtcgttggggtgtgagctatatctatggtacgtgggcagcactgaa
aggtctggcagccgttggtgttccggaaacacatgttagcgtgcagaaaggtattcgtttcctgctgagcaaacagcgtccggatggtggatggggtgaatcatgttatagtgatgcagaagat
cgttttgttccgctgagctttagcacaccggttcagaccgcatgggcattagatgcactgattgcatatcatgatcatcctacaccggcaattgaaaaaggtatggcctgtctgctggaaatgatg
gaaaaacgcggtgaagaatggtcatatccggcaggcgcaggtctgtcaggtcagttttatgtttattatcacagctacccgtatgtttggagcctgatggcaatgacccattatctgcagaaatat
agctaa

27. *Afa*SHC

atgaacaccattagccatccgagcaaagttaaagcagcagttagcgcaaccgttccgcctacaccgagctgtgttaccccgacaccgtttaccggtatgggtaatagcctggcacataccgt
tgcagcagcatgtgattggctgattggtgaacagaaagcagatggtcattgggttggtccggttgcaagcaatgcaagcatggaagcagaatggtgtctggcactgtggtatctgggtttaga
agatcatccgctgcgtccgcgtctgggtaaagcactgctgcacatgcagcgtgaagatggtagctggggcacctattggggtgcaggtaatggtgatattaatgcaaccgttgaagcctatgc
agcactgcgtagcctgggttatgcagcagatacaccggaactgagcaaagcatgtgcatggattatgcgtatgggtggtctgcgtaatgttcgtgtttttacccgttattggctggcactgattggc
gaatggccgtgggaacagaccccgaatctgcctccggaagttatttggtttccgaacaaattcgtgttcagcatctataactttgcacagtgggcacgtgcaaccctggttccgctggcaattct
gagcgcacgtcgtccgagccgtccgttacgtccgcaggatcgtctggatgcactgtttccgcagggtcgtgaaaattttgattatgtgctgccgaaaaaagaaggcgttgatctttggagcagct
tttttcgtaccaccgataaaggtctgcattggctgcagagccgttttctgaaacgtaataccgttcgtgaagcagcaattcgccacatgctggaatggattattcgtcatcaggatgcagatggtg
gttggggtggtattcagcctccgtgggtttatggtctgatggcactgcatggtgaagattatcagtttcatcatccggttatggcaaaagcactggcagcactggatgatcctggttggcgtcgtgat
cagggtgatgcaagctgggttcaagcaaccaatagtccggtttgggataccatgctggccctgatggcattacatgatgcaaatgcagaagaacgttatacaccgcagatggataaagccc
tggactggctgctggcacgtcaggttcgtgttaaaggtgattggagcattaaactgccggatgttgaacctggtggctgggcatttgaatatgccaatgatcgttatcctgataccgatgataccg
cagttgccctgattgcactgagcagctgtcgtaatcgtgaagaatggaaagaaaaaggtgtggaagatgcaattacccgtggtgttaattggttaattgcaatgcagagcagttgtggcggttg
gggagcatttgataaagataataatcgtagcctgctgagcaaaatcccgttttgtgattttggtgaagcactggaccctccgagcgttgatgttaccgcacatgttctggaagcatttggtctgctg
ggtgttccgcgtcagacaccggcactgcaacgtggtctggcatatattcgtgcagaacaagaggcaagcggtgcatggtttggccgttggggtgtgaattatctgtatggtacaggtgcagttct
gcctgcgctggcagcaattggcgaagatatgacccagccgtatattacacgtgcctgcgattggcttattgcacatcagcaagaggatggcggatggggtgaaagctgtgcaagctatatgg
atgttagcagcattggttggggcaccaccacaccgagccagaccgcatgggcactgatgggtctgattgcagcaaatcgtgaacaggaccatccggcaattgcacgtggttgtcgttatctg
attgatcgtcaagaaaccgatggtagttggaccgaagaagaattcaccggcaccggttttccaggttatggtgttggtcagacaattaaactggacgatccagcagttgcaaaacgtctgcag
cagggtgcagaactgagccgtgcatttatgctgcgttatgatctgtatcgtcagtttttccgctgatggccctgagtcgtgcagcacgtattatgccggttggtcagtaa

28. *Aor*SHC

atgaccacaccgctgtttaaaggtatgggtaatagtctgacccataccgttagcagcgcatgtgaatggctgattagccagcagaatccggatggtcattgggttggtccggttggtagcaatg
caagcatggaagcagaatggtgtctggcactgtggtttctgggtttagaagatcatccgctgcgtccgcgtctgggtaatgcactgctgcagacccagcgtgaagatggtagctgggatgtttat
ttaggtgcaggtaatggtgatattaatgcaaccgttgaagcctatgcagcactgcgtagcctgggttatccggaaaatacaccggcactgcagaaagcagcaacctggattaaacagaaag
gtggcctgaaaaacattcgtgtgtttacccgttattggctggcactgattggtgaatggccgtgggaaaaaaccccgaatctgcctccggaaattatctggtttccgaacaaattcgtgttcagca
tctataactttgcacagtgggcacgtgcaaccctggttccgctggcaattctgagcgcacgtcgtccgagccgtccgttacgtccgcaggatcgtctgaatgcactgtttccggaaggtcgtggt
aattttgattataccctgccgaaaaaagaaggcgttgatctttggagcgatttttttcgtaccaccgataaaggtctgcattggctgcagagcaaatttctgaaacgtaataccatgcgtgaagca
gcaattcgtcacatgctggaatggattattcgtcatcaggatgcagatggtggttggggtggtattcagcctccgtgggtttatggtctgatggccctgcatggtgaaggttatcagtttcatcatccg
gttatggcaaaaggtctggatgccctgaatgatcctggttggcgtcatgataaaggtaatgccagctggattcaggcaaccaatagtccggtttgggataccatgctggccattatggcactgc
atgatgcaaaagcagaagatcgttttacaccgcaggttgataaagcattaggttggctgctggatcgtcaggttcgtgttaaaggtgattggagcattaaactgccggatgttgaacctggtggc
tgggcatttgaatatgccaatgatttttatccggacaccgatgataccgcagttgccctgattgcactggcaagctgtcgtcatcgtccggaatggcaagaacgtggtgttgaagatgcaattgc
ccgtgcagttcgttggctggttgccatgcagagcagctgtggcggttggggagcatttgataaagataataacaaagccctgctgagcaaaatcccgttttgtgattttggtgaagcactggacc
ctccgagcgttgatgttaccgcacatattctggaagcatttggtctgctgggcctgcctcgtgatctgccgtgtattaaacatgcactggattatgttcgtgcagaacaggacccgcaaggtccgt
ggtttggccgttggggtgttaattatgtttatggtacaggtgcagttctgccagcactggcagcaattggtgaagatatgacccagccgtatattaccaaagcctgtgattggctggtagcacatca

gcaagaggatggcggatggggtgaaagctgtgcaagctatatggatgcaagcaccattggtcgcggtaaaaccacaccgagccagaccgcatgggcactgatgggtctgattgcagca
gcacgtcctcaggattatccggcaattgaaaaaggttgtcgctatctgattgatcgtcaagaaccggatggcagctggaccgaagaagattacaccggtacaggttttccaggttatggtgttg
gtcagaccattcgtctggatgatccggcactgagcaaacgcctgcagcagggtgcagaactgagccgtgcatttatgctgcgttatgatctgtatcgtcagtttttccgattatggccctgagtcg
tgcaagccgtctgattagtccggaaaccgcaaccgaacaggcagttgaagcagccgcaaaaaatctggaaaagattattgcctaa

29. *Gfr*SHC

atgggtgtttggcgtatgagcgcaccgattctgaaaggtatgagcaatagcctggcacataccgttagctgtgcatgtgattggctgattggtcagcagaaagcagatggtcattgggttggtag
cgttgaaagcaatgcaagcatggaagcagaatggtgtctggcactgtggtttctgggtttagaagatcatccgctgcgtccgcgtctgggtaatgcactgctggaaatgcagcgtgatgatggt
gcatggggtgtttatctgggtgcacagagcggtgatattaatgcaaccgttgaagcctatgcagcactgcgtagcctgggttatagcgcaaatagtccggttctgctgaaagccggtgcatgga
ttagcgaaaaaggtggtctgaaaaacattcgtgtgtttacccgttattggctggcactgattggtgaatggccgtgggaaaaaaccccgaatctgcctccggaaattatctggtttccgaacaatt
ttgtgttcagcatctataactttgcacagtgggcacgtgcaaccctggcaccgctggcaattctgagcgcacgtcgtccgagccgtccgttacgtccgcaggatcgtctggatgcactgtttccg
gaaggtcgtgaaaaatttgattataccctgccgaaaaaagatcgcgttgatctgtggtctagctttttcgtaccaccgataaaggtctgcattggctgcagagccgttttctgaaacgtaataccg
ttcgtgaagcagcaattcgtcacatgctggaatggattattcgtcatcaggatgcagatggtggttggggtggcattcagcctccgtgggtttatggtctgatggccctgcatggtgaaggttatcc
gtttcatcatccggttatggcaaaagcactggcagcactggatgatcctggttggcgttatgatcgtggtgaagcaagctggattcaggcaaccaattcaccggtttgggataccatgctggccc
tgatggcattacatgatgcaaatgcacaagaacgttttacaccggaaatggataaagcattaggttggctgctggaacgtcaggttcgtgttaaaggtgattggagcattaaactgccggatgtt
gaaccaggtggttggagctttgaatatgcaaatgatcgttatccggacaccgatgataccgcagttgccctgattgcactgagcttttgtcgtcatcgtgaagaatggaaacagaaaggcgttg
ataaagcaattgatcgtgcagtgaactggctgatcgcaatgcagagcagctgtggtggctggggagcatttgataaagataataacaaaagcctgctgagcaaaatcccgttttgtgattttgg
tgaagcgctggaccctccgagcgttgatgttaccgcacatattctggaagcatttggtctgctgggtctgagccgtgatctgccggttgttcagaaagccctggcctatgttcgtctggaacagga
cccgcaaggtccgtggtttggccgttggggtgttaattatctgtatggtacaggtgcagttctgcctgcgctggcagcaattggtgaagatatgacccagccgtatatcctgaaagcgtgcgaat
ggctgattagctgtcagcaggatgatggcggatggggagaaagctgtgcaagctatatggatattagcagcattggtcgtggtagcaccaccgcaagccagaccgcatgggcactgatgg
gtctgattgcagttggtcgtcctcaggatcatgaagcaattgcaaaaggttgtcgctttctgattgatcgtcaagaggcagatggtagctggaccgaagaagaattcacaggcaccggttttcca
ggttatggtgttggtcagacaattaaactggacgatccggcactgagcaaacgtctgatgcagggtgcagaactgagtcgtgcatttatgctgcgctatgatatgtatcgtcagtattttccgatta
tggcactggcacgtgcgagccgtctgctgacacaggatatttaa

30. *Kna*SHC

atgaatagcgaaagccgtctgagccgtaaacaggcaggcgcaccgggtcctgataaaattgaagcacgtccggatagcacaccggcagcagcatttcgtggtattgataatagcctgaca
cataccctgagcagcgcatgtaaatggctgatggaacagcagaaaccggatggtcattgggttggtagcgttgcaagcaatgcaagcatggaagcagaatggtgtctggcactgtggtttct
gggtttagaagatcatccgctgcgtccgcgtctgggtaaagcactgctggaaatgcagcgtgaagatggtagctggggtatctattatggtgcaggtaatggtgatattaatgccaccgttgaa
agctatgcagcactgcgtagcctgggttatgcagcagatgatccggcactgagccgtgcagcaacctggattgccagcaaaggtggtctgcgtaatgttcgtgtttttacccgttattggctggc
actgattggtgaatggccgtgggaaaaaaccccgaatctgcctccggaaattatctggtttccgaacaaattcgtgttcagcatctataactttgcacagtgggcacgtgcaaccctggttccgc
tggcaattctgagcgcacgtcgtagcagccgtccgttacgtccgcaggatcgtctggatgccctgtttcctggtggtcgtgaaaattttgattatgaactgcctccgcgtgatggtcaggatctgtg
ggcaacctttttcgtaaaaccgatcgtgcactgcattggctgcagaccaaatttctgaaacgtaataccatgcgtgaagcagcaattcgtcacatgctggaatggattattcgtcatcaggatgc
agatggtggttggggtggtattcagcctccgtgggtttatggtctgatggccctgcatggtgaagattatcagtttcatcatccggttatggcaaaaggtctggcagcactggatgatcctggttgg
cgttatgatcgtggtgatgcaagctgggttcaagcaaccaatagtccggtttgggataccatgctggccctgatggcattacatgatgccgatgccgaaaccgattttacaccggaaatggata
aagcattaggttggctgctggaacgtcaggttcgtgttaaaggtgattggagcgttaaactgccggatctggaacctggtggctgggcatttgaatatgccaatgatcgttatcctgataccgatg
ataccgcagttgccctgattgcactggcagcatgtcgtgatcgtgaagaatggaaaggtcgtggtgttgaagccgcaattacccgtggtgttaattggctggttggtatgcagagcacctgtggc
ggttggggagcatttgataaagataataatcgtgccctgctgagcaaaattccgttttgtgattttggtgaagcactggaccctccgagcgttgatgttaccgcacatgttctggaagcatttggtgt
tctgggtctgcctcgtgatatgcctgcactgcagcgtggtctggcatatattcgtgcagaacaagaggcagatggtccgtggtttggtcgctggggtgtgaattatctgtatggtacaggtgcagtt
ctgcctgcgctggcagcaattggcgaagatatgacccagccgtatattgcccgtgcatgtgattggttagttgcacatcagcaagaaaatggcggatggggtgaaagctgtgcaagctatatg
gaaattgcaagcattggtcgtggtccgaccacaccgagccagaccgcatgggcactgatgggtctgattgcagcaaatcgcaaacaggatcatgaagcaattgttcgtggttgccgttatctg
attgatcagcagcaggccgatggtagttgggaagaaaaagaattcaccggcaccggttttccaggttatggtgttggtcagacaattaaactggacgatccagcactgaccagccgtctgca
gcagggtgcagaactgagtcgtgcatttatgctgcgctatgatctgtatcgtcagtttttccgattatggcactgtcacgtgcagttcgtgttctgaaaggtagcaaataa

31. *Kxy*SHC1

atggataccaccggtcatacaccggttaccaccgcaccggcagcaccggatgcaaccggtacacagacccagaccgcaggcgcaccgtttgcaggtatgggtaattctatgacccatac
cattagcgcagcatgtgattggctgattcagcagcagaaaccggatggtcattgggttggtagcgtgggtagcaatgcaagcatggaagcagaatggtgtattgcactgtggtttctgggtttag
aaaatcatccgctgcgtccgcgtctgggtaatgcactgctggaaatgcagcgtgaagatggtagctggggtgtttatcatggtgcaggtaatggtgatattaatgcaaccgttgaagcctatgc
agcactgcgtagcctgggttatccggcagatacaccggcactggcacgtgcagcaacctggattgcacgtaaaggtggtctgcgtaatattcgtgtttttacccgttattggctggcactgattgg
tgaatggccgtgggaaaaaaccccgaatctgcctccggaaattatctggtttccgaataaattcgtgttcagcatctataactttgcacagtgggcacgtgcgaccctggttccgctggcaattct
gagcgcacgtcgtccgagccgtccgttacgtccgcaggatcgtctggatgcactgtttccggaaggtcgtgcaaattttgattataccctgccgaaaaaagaaggtcgcgatctttgggcaac
ctttttcgtaccaccgatcgtggtctgcattggctgcagagcaatgttctgcgtcgtagcaccatgcgtggtgcagcaattcgccacatgctggaatggattattcgtcatcaggatgcagatggt
ggttggggtggtattcagcctccgtgggtttatggtctgatggccctgcatggtgaagattatcagctgcatcatccggttatggcacgttcactgggtgcactggatgatcctggttggcgtcatga
tcgtggtaatgccagctggattcaggcaaccaatagtccggtttgggataccatgctggccctgatggcattacatgatgcaggcggtgaagatcgttttacaccggaaatggatcgtgccctg
gactggctgctggcacgtcaggttcgtgttcgtggtgattggagcattaaactgccggatgttgaacctggtggctgggcatttgaatatgccaatgatcgttatcctgataccgatgataccgca
gttgccctgattgccctggcaccgtgtcgtaatcgtccggaatggaaagaaaaaggtgttgatgcagccattgatcgtgcagttcgttggctggttgcaatgcagagcgaatgcggtggatggg

gtgcatttgataaagataataatcgtagcctgctgagcaaaatcccgttttgtgattttggtgaagcactggaccctccgagcgttgatgttaccgcacatattctggaagcatttggtctgctgggtctgcctcgtgatatgcctgcaattcagcgtgcactggcctatgttcgtgcagaacaagatccggcaggtccgtggtttggccgttggggtgttaattatgtttatggtacaggtgcagttctgcctgcactggcagcaattggcgaagatatgacccagccgtatattgcccgtgcctgcgattggctggtagcacatcagcaagaggatggcggatggggagaaagctgtgcaagctatatggaaattgcaagcgttggtcgtggcaccaccacaccgagccagaccgcctgggcagttatgggtttagttgcagcaaatcgtgcacaggattatccagcaattgcgcgtggttgtcgttatctgattgaacgtcagcagccggatggtagttggcatgaagccgaattcaccggcaccggttttccaggttatggtgttggtcagacaattaaactggatgacccgatgctgagccagcgtctgagccagggtgcagaactgagccgtgcatttatgctgcgttatgatctgtatcgtcagctgtttccgattatggccctgagtcgtgcagcacgtctgatgccggttggtggcgcaaaacagcagggtacagtttaa

32. *Kxy*SHC2

atgacccgtgaaagccgtccgctgaccaaaaccgcagcaagcagcggcaatattacaagcggtaatacctttccggcagttggcacccagagcgttaatggtggtggtaaaagcaccggtagcgcaagcgcactgcgtacaatggataatagcctgagccatgcaattagcagcgcatgtgattggctggttggtcagcagaaaccggatggtcattgggttggtccggttgcaagcaatgcaagcatggaagcagaatggtgtctggcactgtggtttctgggtttagatgatcacccgctgcgtccgcgtctgggtaaagcactgctggaaatgcagcgtgaagatggtagctggggcacctattatggtgcaggtaatggtgatattaatgccaccgttgaaagctatgcagccctgcgtagcctgggttatccggcagatgatccggcaattagccgtgcagcaacctggattgccagcaaaggtggtctgaaaaacattcgtgtttttacccgttattggctggcactgattggtgaatggccgtgggaaaaaaccccgaatctgcctccggaagttatttggtttccgaacaattttgtgttcagcatctataactttgcacagtgggcacgtgcaaccctggttccgctggcaattctgagcgcacgtcgtccgagtcgtccgttacgtccgcaggatcgtctggatgccctgtttcctggtggtcgtgcaaattttgattatgaactgcctgcacgtggtgatcgtgatctgtgggatcgtttttttcgtgcaaccgatcgtggtctgcattggctgcagagccgttttctgaaacgtaataccctgcgtgaagcagcaattcgtcacatgctggaatggattattcgtcatcaggatgcagatggtggttggggtggtattcagcctccgtgggtttatggtctgatggccctgcatggtgaagattatcagtttcatcatccggttatggcaaaagcactgagtgcactgaatgatcctggttggcgtcatgataaaggtgatgcaagctggattcaggcaaccaatagtccggtttgggataccatgctggccattatggcactgcatgatgccgatggtgaaacccagtttagtccgcagatggaaaaagcattaggttggctgctggatcgtcaggttcgtgtgaaaggtgattggagcattaaactgcctgatgttgaacctggtggctgggcatttgaatatgccaatgatcgttatcctgataccgatgataccgcagttgccctgattgcactgagcagctgtcgtaatcgtgaagaatggaaaaaacgtggtgttgaagaggcaatttctcgtggtgttaattggctgattggcatgcagtcagaatgtggcggttggggagcatttgataaagataataatcgtagcatcctgagcaaaatcccgttttgtgattttggtgaagcactggaccctccgagcgttgatgttaccgcacatgttctggaagcatttggtattctgggtctgcctcgccacatgccgaccattcagcgtgcactggcatatattcgtgcagaacaagaacctgatggtccgtggtttggtcgctggggtgtgaattatctgtatggtacaggtgcagttctgccagcactggcagcaattggcgaagatatgacccagccgtatattaccaaagcctgcgattggcttgttgcacatcagcaagaaaatggcggatggggtgaaagctgtgcaagctatatggaactgagcatggttggtcgtggtgtgaccacaccgagccagaccgcatgggcactgatgggtctgattgcagcaaatcgtcctcaggattatggcgcaattgcccgtggttgtcgttatctgattgatctgcagcaggcagatggttcatggcatgaaaaagaattcaccggcaccggttttccaggttatggtgttggtcagacaattaaactggatgatccagcgctgagcaaacgtctgcagcaaggtgcagaactgagccgtgcatttatgctgcgttatgatctgtatcgtcagtttttccgattatggccctgagtcgtgcaagccgtctgatgaaattagaaaaataa

33. *Aci*SHC_R1.1

atgacccaggcaagcgttcgtgaagatgcaaaagcagcactggatcgtgcagttgattatctgctgagcctgcaggatgaaaaaggtttttggaaaggtgaactggaaaccaacgttaccattgaagccgaagatctgctgctgcgtgaatttctgggtattcgtacaccggatattaccgcagaaaccgcacgttggattcgtgcaaaacagcgtagtgatggcacctgggcaaccttttatgatggtccgcctgatctgagcaccagcgttgaagcctatgttgcactgaaactggctggtgatgatccggcagcaccgcacatggaaaaagcagccgcatatattcgtggtgccggtggtgttgaacgtacccgtgtttttacccgtctgtggctggcactgtttggtctgtggccgtgggatgatctgccgacactgcctccggaaatgatttttctgccgagctggtttccgctgaacatttatgattggggttgttgggcacgtcagaccgttgtgccgctgaccattgttagcgcactgcgtccggttcgtccgattccgctgagtattgatgaaattcgtacaggtgcaccgcctccgcctcgcgatccggcatggaccattcgtggtttttttcagcgtctggatgatttactgcgtggttatcgtcgtgttgcagatcatggtccggcacgtctgtttcgtcgtctggcaatgcgtcgtgcagcagaatggattattgcacgtcaagaggcagatggtagctggggtggtattcagcctccgtgggtttatagcctgattgcactgcatctgctgggttatccgctggatcatccggttctgcgtcgtggtctggatggtctgaatggttttaccattcgcgaagaaacagcagatggtgcagttcgtcgcctggaattttgccagagtccggtttgggataccgcactggcagttacagcactgcgtgatgcaggtctgcctgccgatcatccgcgtgttcaggcagcagcccgttggctggttggtgaagaggtgcgtgttgccggtgattgggcagtgcgtcgtccgggtctgcctcctggtggttgggcatttgaatttgccaatgataattacccggataccgatgatacagcagaagttgttctggccctgcgtcgcgttcgtctggaagatgcagatcagcaggcgctggaagcagccgttcgtcgtgcaaccacctgggttattggtatgcagagcaccgatggcggttggggtgcatttgatgcagataatacccgtgaactggtgctgcgtctgccgttttgtgattttggtgccgttattgatccgcctagcgcagatgttaccgcacatattgttgaaatgctggcagccctgggtatgcgtgatcatcctgccaccgttgcgggtgttcgttggctgctggcacatcaagaacctgatggtagttggtttggtcgttggggagcaaatcatatttatggtacgggtgcagttgttccggcactgattgcagccggtgttagtccggatacaccgcctattcgtcgcgcaattcgctggctggaagaacatcagaatccggatggtggatggggtgaagatttacgtagctataccgatcctgcactgtgggttggtcgtggtgttagcaccgcaagccagaccgcatgggcactgctggcattactggcagcgggtgaagaagcaagtccggcagttgatcgtggcgttcggtggctggttaccacacagcagcctgatggtggctgggatgagccgcattacaccggtacaggttttccgggtgatttctatattaactatcatctgtatcgcctggtgtttccgattagtgcactgggtcgttatgttaatcgt

34. *Aci*SHC_R2.1

atgacccaggcaagcgttcgtgaagatgcaaaagcagcactggatcgtgcagttgattatctgctgagcctgcaggatgaaaaaggtttttggaaaggtgaactggaaaccaacgttaccattgaagccgaagatctgctgctgcgtgaatttctgggtattcgtacaccggatattaccgcagaaaccgcacgttggattcgtgcaaaacagcgtagtgatggcacctgggcaaccttttatgatggtccgcctgatctgagcaccagcgttgaagcctatgttgcactgaaactggctggtgatgatccggcagcaccgcacatggaaaaagcagccgcatatattcgtggtgccggtggtgttgaacgtacccgtgtttttacccgtctgtggctggcactgtttggtctgtggccgtgggatgatctgccgacactgcctccggaaatgatttttctgccgagctggtttccgctgaacatttatgattggggttgttggccgcgccagaccgttgtgccgctgaccattgttagcgcactgcgtccggttcgtccgattccgctgagtattgatgaaattcgtacaggtgcaccgcctccgcctcgcgatccggcatggaccattcgtggtttttttcagcgtctggatgatttactgcgtggttatcgtcgtgttgcagatcatggtccggcacgtctgtttcgtcgtctggcaatgcgtcgtgcagcagaatggattattgcacgtcaagaggcagatggtagctggggtggtattcagcctccatgggtttatagcctgattgcactgcatctgctgggttatccgctggatcatccggttctgcgtcgtggtctggatggtctgaatggttttaccattcgcgaagaaacagcagatggtgcagttcgtcgcctggaaatgtgccagagtccggtttgggataccgcactggcagttacagcactgcgtgatgcaggtctgcctgccgatcatccgcgtgttcaggcagcagcccgttggctggttggtgaagaggtgcgtgttgccggtgattgggcagtgcgtcgtccgggtctgcctcctggtggttgggcatttgaatttgccaatgataattacccgg

ataccgatgatacagcagaagttgttctggccctgcgtcgcgttcgtctggaagatgcagatcagcaggcgctggaagcagccgttcgtcgtgcaaccacctgggttattggtatgcagagca
ccgatggcggttggggtgcatttgatgcagataatacccgtgaactggtgctgcgtctgccgttttgtgattttggtgccgttattgatccgcctagcgcagatgttaccgcacatattgttgaaatgc
tggcagccctgggtatgcgtgatcatcctgccaccgttgcgggtgttcgttggctgctggcacatcaagaacctgatggtagttggtttggtcgttggggagcaaatcatatttatggtacgggtgc
agttgttccggcactgattgcagccggtgttagtccggatacaccgcctattcgtcgcgcaattcgctggctggaagaacatcagaatccggatggtggatggggtgaagatttacgtagctat
accgatcctgcactgtgggttggtcgtggtgttagcaccgcaagccagaccgcatgggcactgctggcattactggcagcgggtgaagaagcaagtccggcagttgatcgtggcgttcggtg
gctggttaccacacagcagcctgatggtggctgggatgagccgcattacaccggtacaggcttcccgggtgatttctatgttaattatcatctgtatcgcctggtgtttccgattagtgcactgggtc
gttatgttaatcgt

35. *Aci*SHC_R2.2

atgacccaggcaagcgttcgtgaagatgcaaaagcagcactggatcgtgcagttgattatctgctgagcctgcaggatgaaaaaggtttttggaaaggtgaactggaaaccaacgttacca
ttgaagccgaagatctgctgctgcgtgaatttctgggtattcgtacaccggatattaccgcagaaaccgcacgttggattcgtgcaaaacagcgtagtgatggcacctgggcaaccttttatgat
ggtccgcctgatctgagcaccagcgttgaagcctatgttgcactgaaactggctggtgatgatccggcagcaccgcacatggaaaaagcagccgcatatattcgtggtgccggtggtgttga
acgtacccgtgtttttacccgtctgtggctggcactgtttggtctgtggccgtgggatgatctgccgacactgcctccggaaatgatttttctgccgagctggtttccgctgaacatttatgattggggtt
gttggccgcgccagaccgttgtgccgctgaccattgttagcgcactgcgtccggttcgtccgattccgctgagtattgatgaaattcgtacaggtgcaccgcctccgcctcgcgatccggcatgg
accattcgtggtttttttcagcgtctggatgatttactgcgtggttatcgtcgtgttgcagatcatggtccggcacgtctgtttcgtcgtctggcaatgcgtcgtgcagcagaatggattattgcacgtca
agaggcagatggtagctggggtggtattcagtggccatgggtttatagcctgattgcactgcatctgctgggttatccgctggatcatccggttctgcgtcgtggtctggatggtctgaatggttttac
cattcgcgaagaaacagcagatggtgcagttcgtcgcctggaaatgtgccagagtccggtttgggataccgcactggcagttacagcactgcgtgatgcaggtctgcctgccgatcatccgc
gtgttcaggcagcagcccgttggctggttggtgaagaggtgcgtgttgccggtgattgggcagtgcgtcgtccgggtctgcctcctggtggttgggcatttgaatttgccaatgataattacccgg
ataccgatgatacagcagaagttgttctggccctgcgtcgcgttcgtctggaagatgcagatcagcaggcgctggaagcagccgttcgtcgtgcaaccacctgggttattggtatgcagagca
ccgatggcggttggggtgcatttgatgcagataatacccgtgaactggtgctgcgtctgccgttttgtgattttggtgccgttattgatccgcctagcgcagatgttaccgcacatattgttgaaatgc
tggcagccctgggtatgcgtgatcatcctgccaccgttgcgggtgttcgttggctgctggcacatcaagaacctgatggtagttggtttggtcgttggggagcaaatcatatttatggtacgggtgc
agttgttccggcactgattgcagccggtgttagtccggatacaccgcctattcgtcgcgcaattcgctggctggaagaacatcagaatccggatggtggatggggtgaagatttacgtagctat
accgatcctgcactgtgggttggtcgtggtgttagcaccgcaagccagaccgcatgggcactgctggcattactggcagcgggtgaagaagcaagtccggcagttgatcgtggcgttcggtg
gctggttaccacacagcagcctgatggtggctgggatgagccgcattacaccggtacaggcttcccgggtgatttctatcttaattatcatctgtatcgcctggtgtttccgattagtgcactgggtc
gttatgttaatcgt

36. *Aci*SHC_R2.3

atgacccaggcaagcgttcgtgaagatgcaaaagcagcactggatcgtgcagttgattatctgctgagcctgcaggatgaaaaaggtttttggaaaggtgaactggaaaccaacgttacca
ttgaagccgaagatctgctgctgcgtgaatttctgggtattcgtacaccggatattaccgcagaaaccgcacgttggattcgtgcaaaacagcgtagtgatggcacctgggcaaccttttatgat
ggtccgcctgatctgagcaccagcgttgaagcctatgttgcactgaaactggctggtgatgatccggcagcaccgcacatggaaaaagcagccgcatatattcgtggtgccggtggtgttga
acgtacccgtgtttttacccgtctgtggctggcactgtttggtctgtggccgtgggatgatctgccgacactgcctccggaaatgatttttctgccgagctggtttccgctgaacatttatgattggggtt
gttggccgcgccagaccgttgtgccgctgaccattgttagcgcactgcgtccggttcgtccgattccgctgagtattgatgaaattcgtacaggtgcaccgcctccgcctcgcgatccggcatgg
accattcgtggtttttttcagcgtctggatgatttactgcgtggttatcgtcgtgttgcagatcatggtccggcacgtctgtttcgtcgtctggcaatgcgtcgtgcagcagaatggattattgcacgtca
agaggcagatggtagctggggtggtattcagtggccatgggtttatagcctgattgcactgcatctgctgggttatccgctggatcatccggttctgcgtcgtggtctggatggtctgaatggttttac
cattcgcgaagaaacagcagatggtgcagttcgtcgcctggaactgtgccagagtccggtttgggataccgcactggcagttacagcactgcgtgatgcaggtctgcctgccgatcatccgc
gtgttcaggcagcagcccgttggctggttggtgaagaggtgcgtgttgccggtgattgggcagtgcgtcgtccgggtctgcctcctggtggttgggcatttgaatttgccaatgataattacccgg
ataccgatgatacagcagaagttgttctggccctgcgtcgcgttcgtctggaagatgcagatcagcaggcgctggaagcagccgttcgtcgtgcaaccacctgggttattggtatgcagagca
ccgatggcggttggggtgcatttgatgcagataatacccgtgaactggtgctgcgtctgccgttttgtgattttggtgccgttattgatccgcctagcgcagatgttaccgcacatattgttgaaatgc
tggcagccctgggtatgcgtgatcatcctgccaccgttgcgggtgttcgttggctgctggcacatcaagaacctgatggtagttggtttggtcgttggggagcaaatcatatttatggtacgggtgc
agttgttccggcactgattgcagccggtgttagtccggatacaccgcctattcgtcgcgcaattcgctggctggaagaacatcagaatccggatggtggatggggtgaagatttacgtagctat
accgatcctgcactgtgggttggtcgtggtgttagcaccgcaagccagaccgcatgggcactgctggcattactggcagcgggtgaagaagcaagtccggcagttgatcgtggcgttcggtg
gctggttaccacacagcagcctgatggtggctgggatgagccgcattacaccggtacaggcttcccgggtgatttctatgttaattatcatctgtatcgcctggtgtttccgattagtgcactgggtc
gttatgttaatcgt

G. References

[1] F. Sievers, A. Wilm, D. Dineen, T. J. Gibson, K. Karplus, W. Li, R. Lopez, H. McWilliam, M. Remmert, J. Söding, J. D. Thompson, D. G. Higgins, *Mol. Syst. Biol.* **2011**, *7*, DOI 10.1038/msb.2011.75.
[2] M. N. Price, P. S. Dehal, A. P. Arkin, *PLoS One* **2010**, *5*, DOI 10.1371/journal.pone.0009490.
[3] T. Frickey, E. Kannenberg, *Environ. Microbiol.* **2009**, *11*, 1224–1241.
[4] D. Ochs, C. Kaletta, K. D. Entian, A. Beck-Sickinger, K. Poralla, *J. Bacteriol.* **1992**, *174*, 298–302.
[5] G. Siedenburg, D. Jendrossek, M. Breuer, B. Juhl, J. Pleiss, M. Seitz, J. Klebensberger, B. Hauer, *Appl. Environ. Microbiol.* **2012**, *78*, 1055–1062.
[6] G. Siedenburg, M. Breuer, D. Jendrossek, *Appl. Microbiol. Biotechnol.* **2013**, *97*, 1571–1580.
[7] M. Perzl, P. Müller, K. Poralla, E. L. Kannenberg, *Microbiology* **1997**, *143*, 1235–1242.
[8] A. Tippelt, L. Jahnke, K. Poralla, *Biochim. Biophys. Acta - Lipids Lipid Metab.* **1998**, *1391*, 223–232.
[9] J. Shinozaki, M. Shibuya, K. Masuda, Y. Ebizuka, *FEBS Lett.* **2008**, *582*, 310–318.
[10] J. Shinozaki, M. Shibuya, Y. Takahata, K. Masuda, Y. Ebizuka, *ChemBioChem* **2010**, *11*, 426–433.
[11] T. Sato, S. Yoshida, H. Hoshino, M. Tanno, M. Nakajima, T. Hoshino, *J. Am. Chem. Soc.* **2011**, *133*, 9734–9737.

[12] T. Sato, H. Hoshino, S. Yoshida, M. Nakajima, T. Hoshino, *J. Am. Chem. Soc.* **2011**, *133*, 17540–17543.

[13] A. Waterhouse, M. Bertoni, S. Bienert, G. Studer, G. Tauriello, R. Gumienny, F. T. Heer, T. A. P. De Beer, C. Rempfer, L. Bordoli, R. Lepore, T. Schwede, *Nucleic Acids Res.* **2018**, *46*, W296–W303.

[14] O. Trott, A. J. Olson, *J. Comput. Chem.* **2009**, *31*, 455–461.

[15] E. Eichhorn, E. Locher, S. Guillemer, D. Wahler, L. Fourage, B. Schilling, *Adv. Synth. Catal.* **2018**, *360*, 2339–2351.

[16] B. A. Baker, Ž. V. Bošković, B. H. Lipshutz, *Org. Lett.* **2008**, *10*, 289–292.

[17] E. Brenna, C. Fuganti, S. Serra, P. Kraft, *European J. Org. Chem.* **2002**, 967–978.

[18] C. Fuganti, S. Serra, A. Zenoni, *Helv. Chim. Acta* **2000**, *83*, 2761–2768.

[19] S. Racolta, P. B. Juhl, D. Sirim, J. Pleiss, *Proteins Struct. Funct. Bioinforma.* **2012**, *80*, 2009–2019.

[20] S. C. Hammer, P. O. Syrén, B. Hauer, *ChemistrySelect* **2016**, *1*, 3589–3593.

[21] A. Schneider, P. Jegl, B. Hauer, *Angew. Chem. Int. Ed.* **2021**, 60, 13251–13256; *Angew. Chem*. **2021**, 133, 13359–13365.

Article VI

1. Introduction:

Studying the recognition and interaction profile between (two) molecules is a common denominator of many life sciences [1]. As such, the prediction of these interactions through automated docking software has wide-ranging applications, from structure-based drug design [2] to the pre-selection of sites in rational design [3], the early stages of drug discovery, and many others. Among the available predictive tools, AutoDock Vina [4] has established itself as one of the most popular and widely used docking engines within the scientific community, fueled by its open-source nature, speed, and ease of use. The software recently received an update that expands its feature set to include additional scoring functions, the simultaneous docking of multiple ligands, flexible residue docking, and a hydrated docking protocol [5]. However, these exciting new features also raise the barrier of entry, as additional steps are required to set up, carry out and analyze the docking protocols.

For this reason, we developed AlphaDock, an open-source PyMOL (Schrödinger, LLC) plugin, to simplify the process of installing, pre-processing, and evaluating docking results in an intuitive graphical user interface (GUI).

2. Overview of Features:

i) Running AutoDock Vina requires the AutoDock Vina binaries [5], several tools from the ADFR software suite [6], and the python package Meeko. To simplify the installation process and avoid version mismatches, we provide a docker container that already contains all requirements and can be set up on any system with a single command.

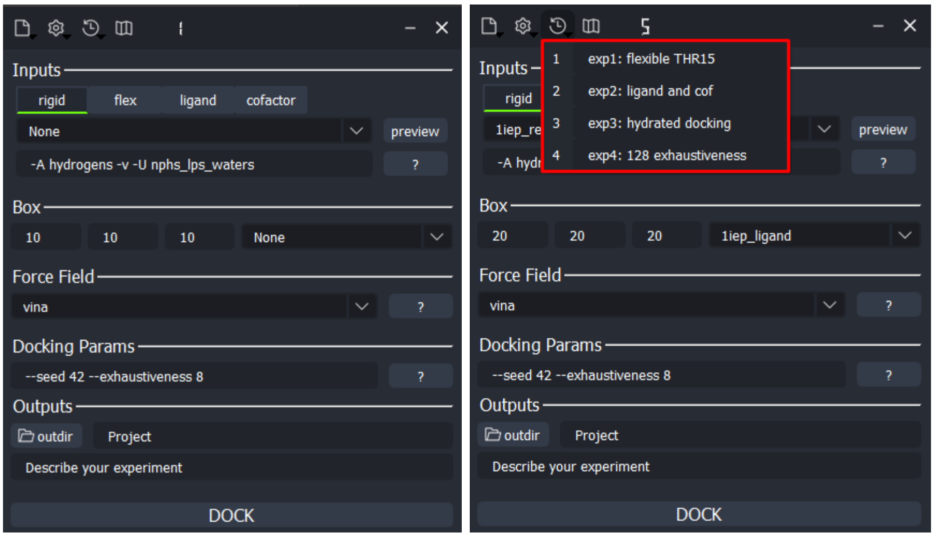
ii) The AlphaDock GUI connects to the running docker container through an SSH connection and can be configured to run locally or remotely anywhere on the network. This makes it possible for researchers to offload expensive computations to a more capable workstation machine. On top of that, running the docking protocols remotely makes the setup for individual users even more accessible, as docker and its dependencies are not required on a user's personal computer.

iii) The AlphaDock GUI provides direct access to AutoDock-Vina v1.2.3, different scoring functions, hydrated docking, multiple ligand docking, and flexible residue docking. All the options and parameters exposed by AutoDock-Vina can be configured through the AlphaDock GUI. We perform the necessary pre-processing from user-selected PyMOL objects and selections.

iv) To avoid reproducibility issues, a detailed history of all input, intermediate, and result files, as well as standard output and standard error, is kept. Users can browse their experiment history and restore/visualize the output of previous runs. On top of the standardized docker container, this history should make all docking experiments precisely reproducible and traceable.

3. Implementation

AlphaDock is a plugin written in Python 3 for PyMOL versions 2.5 and higher. The docking engine and its dependencies are packaged in a docker container and, as such, are compatible with any system that can run docker. However, docker is only required when docking locally. Additional information on how to install and configure AlphaDock can be found on the AlphaDock GitHub page: https://github.com/ccbiozhaw/dock.



None

-A hydrogens -v -U nphs_lps_waters

10 10 10 None

vina

--seed 42 --exhaustiveness 8

1 exp1: flexible THR15
2 exp2: ligand and cof
3 exp3: hydrated docking
4 exp4: 128 exhaustiveness

1iep_re

-A hydr

20 20 20 1iep_ligand

vina

--seed 42 --exhaustiveness 8

Figure 1: AlphaDock GUI. Dropdown menus list all available PyMOL *selections and objects in the inputs section. All features and configuration options the underlying tools provide can be accessed and changed. The "?" buttons display additional information. All experiments are tracked and displayed in the experiment history.*

[1] A. Grosdidier, V. Zoete, and O. Michielin, "SwissDock, a protein-small molecule docking web service based on EADock DSS," *Nucleic Acids Res*, vol. 39, no. SUPPL. 2, Jul. 2011, doi: 10.1093/nar/gkr366.

[2] V. Y. Tanchuk, V. O. Tanin, A. I. Vovk, and G. Poda, "A New, Improved Hybrid Scoring Function for Molecular Docking and Scoring Based on AutoDock and AutoDock Vina," *Chem Biol Drug Des*, vol. 87, no. 4, pp. 618–625, Apr. 2016, doi: 10.1111/cbdd.12697.

[3] M. T. Reetz, L. W. Wang, and M. Bocola, "Directed evolution of enantioselective enzymes: Iterative cycles of CASTing for probing protein-sequence space," *Angewandte Chemie - International Edition*, vol. 45, no. 8, pp. 1236–1241, Feb. 2006, doi: 10.1002/anie.200502746.

[4] O. Trott and A. J. Olson, "AutoDock Vina: Improving the speed and accuracy of docking with a new scoring function, efficient optimization, and multithreading," *J Comput Chem*, p. NA-NA, 2009, doi: 10.1002/jcc.21334.

[5] J. Eberhardt, D. Santos-Martins, A. F. Tillack, and S. Forli, "AutoDock Vina 1.2.0: New Docking Methods, Expanded Force Field, and Python Bindings," *J Chem Inf Model*, vol. 61, no. 8, pp. 3891–3898, Aug. 2021, doi: 10.1021/acs.jcim.1c00203.

[6] S. Forli, R. Huey, M. E. Pique, M. F. Sanner, D. S. Goodsell, and A. J. Olson, "Computational protein-ligand docking and virtual drug screening with the AutoDock suite," *Nat Protoc*, vol. 11, no. 5, pp. 905–919, May 2016, doi: 10.1038/nprot.2016.051.

Eigenständigkeitserklärung

Hiermit erkläre ich, dass diese Arbeit bisher von mir weder an der Mathematisch Naturwissenschaftlichen Fakultät der Universität Greifswald noch einer anderen wissenschaftlichen Einrichtung zum Zwecke der Promotion eingereicht wurde.

Ferner erkläre ich, dass ich diese Arbeit selbstständig verfasst und keine anderen als die darin angegebenen Hilfsmittel und Hilfen benutzt und keine Textabschnitte eines Dritten ohne Kennzeichnung übernommen habe.

David Patsch

Curriculum Vitae

Persönliche Angaben

Vorname:	David
Nachname:	Patsch
Geburtsdatum:	15.11.1992
Staatsangehörigkeit:	Österreich

Ausbildung

04/2020 - heute	**Promotionsstudium** Universität Greifswald / ZHAW Wädenswil
08/2017 – 08/2019	**Master Biotechnology** Management Center Innsbruck (MCI)
10/2012 – 08/2017	**Bachelor Biology** University of Innsbruck
2003 – 2011	**High School** Gymnasium Feldkirch

David Patsch

List of Publications

Eichenberger, M.*, Hüppi, S.*, Patsch, D.*, Aeberli, N., Berweger, R., Dossenbach, S., Eichhorn, E., Flachsmann, F., Hortencio, L., Voirol, F., Vollenweider, S., Bornscheuer, U., & Buller, R. Asymmetric Cation–Olefin Monocyclization by Engineered Squalene–Hopene Cyclases. *Ang. Chem. Int. Ed.*, **2021**, 60(50), 26080–6. DOI: 10.1002/anie.202108037

* equal contribution

J. Büchler, S. Honda Malca, D. Patsch, M. Voss, N. J. Turner, U. T. Bornscheuer, O. Allemann, C. Le Chapelain, A. Lumbroso, O. Loiseleur, R, Buller. Algorithm-aided engineering of aliphatic halogenase WelO5* for the asymmetric late-stage functionalization of soraphens. *Nat. Commun.,* **2022**, 13(1), 371. DOI: 10.1038/s41467-022-27999-1

D. Patsch, R. Buller, Improving Enzyme Fitness with Machine Learning. CHIMIA, **2023**, 77(3), 116. DOI: 10.2533/chimia.2023.116

D. Patsch, M. Eichenberger, M. Voss, U. T. Bornscheuer, R. Buller. LibGENiE – A bioinformatic pipeline for the design of information-enriched enzyme libraries. Submitted to *Comput. Struct. Biotechnol. J.*, **2023**.

Acknowledgments

I wish to express my gratitude to everyone who has contributed to completing my Ph.D. thesis. Your support and guidance have been invaluable throughout this journey.

First and foremost, I would like to extend my deepest appreciation to Rebecca for your supervision and support at every stage of my doctoral research. I genuinely appreciate the freedom I had to pursue different ideas and directions. I suspect this is something unique to a Ph.D. and what I will miss the most moving forward.

I also want to express my gratitude to Prof. Bornscheuer for accepting me as a Ph.D. student.

I was privileged to work in an incredible group filled with amazing people. In particular, I want to thank Michi Eichenberger for his help, discussions, patience, and direction. You heavily influenced a lot of the work in this thesis. Thanks to the people in RT for the fun (scientific) debates and drama, Fabian, Sean, Johannes, Thomas, Athena, Katrin, Nicolas, Michi N. and Michi E.; the corona time could have been much worse. We ended up always laughing, even in new and different situations. I would also like to thank the people in the RU, Moritz, Sumire, Nadine, Daniela M., Gigersan, Eimear, and Daniela S.

To all the members of our research group, I am grateful for your generous support, stimulating discussions, and enjoyable coffee breaks.

I would also like to express my heartfelt appreciation to my family. Thank you to my parents and sister, who have supported me from the beginning and accompanied me with unwavering trust.

Once again, I extend my deepest gratitude to all those mentioned above and to anyone else who has contributed to my academic and personal growth. Your support has been truly invaluable, and I am forever grateful for your presence in my life.