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

Johannes Büchler¹,², Sumire Honda Malca¹, David Patsch¹,³, Moritz Voss¹, Nicholas J. Turner², Uwe T. Bornscheuer³, Oliver Allemann⁴,⁵, Camille Le Chapelain⁴, Alexandre Lumbroso⁴, Olivier Loiseleur⁴,⁵ & Rebecca Buller¹

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 sp³ 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.
Subtle molecular changes in small molecules can have a profound impact on their biological activity and metabolism. For example, monochlorinated and dichlorinated vancomycin lose approximately 30 and 50% of the antimicrobial effect exhibited by parent antibiotic vancomycin, respectively. Similarly, the introduction of a single methyl group led to MK-8133, a dual orexin receptor antagonist, with 480-fold boosted potency. In the latter example, the methyl group had to be installed through a laborious five-step de novo synthesis. 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, 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 malleability of non-natural substrates en route to potency 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 malleability of non-natural substrates en route to potency while helping to improve physical properties such as solubility and stability.

The synthesis route to organohalides commonly involves multiple steps. In order to achieve high chemico-, regio- and stereoselectivities, 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. Enzymatic halogenations, on the other hand, often exhibit excellent regio- and stereoselectivity even in complex molecular settings, therefore complementing—and sometimes outperforming—existing strategies.

Biocatalytic halogenations are carried out by enzymes called halogenases, which are typically classified according to their catalytic mechanism: Heme, vanadium, and flavin-dependent halogenases (Fi-Hals) follow an electrophilic aromatic substitution mechanism via the generation of hypohalous acid, iron/halogenases (Fl-Hals) follow an electrophilic aromatic substitution mechanism, and iron/halogenases WelO5s with the aim to provide catalytic activity against several phytopathogenic fungi through inhibition of the yeast biotin carboxylase (BC) domain with soraphen A (PDB ID: 1W96) revealed that the macrolide acts as an allosteric inhibitor by disrupting dimerization of the BC domain and stabilizing the catalytically inactive monomer (Supplementary Fig. 1). 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. Notably, soraphen A has recently also become a target of pharmaceutical interest. In cancer therapy research, several studies established that tumoral cells have a dependence on fatty acid synthesis and that inhibition of ACC triggers apoptosis with no or little effects on healthy cells. 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. 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 aKGHs 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 highly interesting but chemically very complex molecules. In addition, such polyketide macromolecules are also known to adopt several conformations. While soraphen A can be accessed through an optimized bioprocess, 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. To obtain the compound for our study, we, therefore, developed a concise semisynthesis starting from soraphen A.
(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 fermentation53, 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 literature31,41,42,54 whereas the engineered aKGS consist of WelO5\(^*\) variants which we had previously identified as possessing a broadened substrate scope43.

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 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 aKGS, 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 U22, led to the predominant formation (ca. 50%) of hydroxylated product and 25% of the desired chlorinated product 12-epi-hapalindole E43. Other examples include studies on the carrier-protein-dependent halogenase SryB2, which turned into an effective hydroxylase in response to the length of added C-atoms in its native substrate L-threonine55.

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-MODEL56, 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 effort57. In order to allow for an unbiased library construction, we opted for a \textit{de novo} library synthesis using high-fidelity on-chip solid-phase gene synthesis58. This library construction strategy allowed us to limit library diversity to the theoretical 8000
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 feature 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, 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 $S = (\text{SIM}_{1a} - \text{SIM}_{1b})/(\text{SIM}_{1a} + \text{SIM}_{1b})$. 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 (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).
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 substrate65. 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 study43 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 \( k_{cat} = 1.96 \) 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 soraphen C, 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 phytotoxity. 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 *Monogynhella 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 Deal69 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 product47. 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
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. 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) 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 203 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, 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. 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. 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 protein and panel 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 auto-overnight in 1 mL of LB medium containing 50 μg/mL kanamycin. A single colony of freshly transformed cells was cultured about 0.1 mL of 25 mM HEPES buffer (pH 8.0) containing 1 mg/mL lysozyme, 0.5 μg/mL polymyxin B, 0.01 mg/mL DNase and incubation for a minimum of 40 °C on a shaking incubator at 850 rpm. Reaction was 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 μM NaCl, 40 μM Glucose, and 2 μM GDH/Ec-Fre®. Incubation and work-up was performed in analogy to the aKGH protocol. The biotransformations with the Fl-Hal library were carried out once including the appropriate controls.

Preparative scale biotransformation. WeLoS® SLP variant was used to prepare compoud 1a and 1b and WeLoS® VWS was used to prepare compound 1c. For the preparation of compound 2a the variant WeLoS® VAA was used. Twenty grams of WeLoS® variant cells were resuspended in 100 mL of lysis buffer (50 mM sodium phosphate, pH 8.0) containing 1 mg/mL lysozyme, 0.5 μg/mL polymyxin B, 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 μL of HEPES buffer (pH 8.0) containing 2 mM substrate, 220 μM 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. Reaction was quenched by the addition of 800 μL methanol/water 5:3 mixture to each well and seeded 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 and performing variants (SLP; WVS) were analyzed in triplicates as individual experiments.

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 MS 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%; 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 (WeLoS® GAP) included as a control.

Construction, expression, and purification of His-tagged WeLoS® variants. His-tagged WeLoS® 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′-GGTTGGGGACAAATCCCTCCTAG-3′ (forward) and 5′-GTTTTTGGCTAGGGCGGATCTCGAG-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 20 °C on a shaking incubator at 850 rpm. The cells were pelleted by centrifugation at 4000 × g for 15 min, and the supernatant was discarded. The cell pellet was stored in a –80 °C freezer prior to biotransformation reactions.
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, NaNO3 and NaNO2), 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 kcat (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, methanol at 14% v/w) 1c. 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).

Ligand docking and homology modeling of WelO5* variants. Models of the wild type WelO5* and the WelO5* variants were created using the SWISS-MODEL online server with default parameters. The crystal structure of wild type WelO5* (PDB ID: 5JAR) served as a template for the homology modeling. The docking process was performed using default parameters of Chimera AutoDock Vina 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. CI conversion WelO5* mutant / tot. CI conversion WelO5* GAP whereas tot. CI conversion = (SIMa + SIMa+1) / (SIMa + SIMa+1 + SIMa+2 + SIMa+3). The selectivity label (S) was calculated using the formula S = (SIMa + SIMa+1) / (SIMa + SIMa+1 + SIMa+2 + SIMa+3). Amino acids were represented as a 17-dimensional vector, which was obtained by concatenating the five-dimensional T-scale descriptors and additional information about the amino acid characteristic. 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.26 The model was implemented in the skil-l Barrister python module. We took a similar approach for predictions of selectivity; however, we excluded variants below a better gauge the generalizability of our model, we cross-validated over ten splits, and 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: 5JAR. 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 diclorovancmycin. 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. & Kraska, 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, 623–635 (2020).
5. Bougé, J. & Ritter, T. Late-stage functionalization. Chem. Soc. Rev. 46, 1197–1237 (2017).
6. Wu, S., Snajdrova, R., Moore, J. C., Baldenius, K. & Bornscheuer, U. T. Machine learning approach for predictions of selectivity; however, we excluded variants below a similar approach for predictions of selectivity; however, we excluded variants below a better gauge the generalizability of our model, we cross-validated over ten splits, and 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: 5JAR. 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 diclorovancmycin. 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. & Kraska, 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, 623–635 (2020).
5. Bougé, J. & Ritter, T. Late-stage functionalization. Chem. Soc. Rev. 46, 1197–1237 (2017).
6. Wu, S., Snajdrova, R., Moore, J. C., Baldenius, K. & Bornscheuer, U. T. Machine learning approach for predictions of selectivity; however, we excluded variants below a similar approach for predictions of selectivity; however, we excluded variants below a better gauge the generalizability of our model, we cross-validated over ten splits, and 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: 5JAR. The authors declare that all the data supporting the findings of this work are available within the
24. Hillwig, M. L., Zhu, Q., Iittiamonkul, K. & Liu, X. Discovery of a promiscuous non-ß-iron halogenase in ambiguous alkaloid biogenesis: implication for an evolvable enzyme family for late-stage sulfonation of aliphatic drugs 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(III)-based 2-oxo-glutarate-dependent halogenase in Menispermaceae plants. Nat. Commun. 11, 1867 (2020).

27. Zhao, C. et al. An Fe²⁺- 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., Patallao, 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 regioselectivemly 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. Yoshida, H., High, Y., Yoshikawa, H., China, V. & Ogino, T. Improvement of the stability and activity of the BPO-A1 haloperoxidase from Streptomyces aureofaciens by directed evolution. J. Biotechnol. 192, 248–254 (2015).

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., Grüsschow, S., Cairns, N. & Goss, R. J. M. Gene expression enabling synthetic diversification of natural products: chemogenetic generation of padmacyamin analogs. J. Am. Chem. Soc. 132, 12243–12245 (2010).

37. Rungunphan, W. & O’Connor, S. E. Diversification of monoterpenic 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. Glotai, D. S. et al. A marine viral halogenase that iodinates diverse substrates. Nat. Chem. 11, 1091–1097 (2019).

41. Andorfer, M. C., Park, H. I., 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 halogenases enable regiocomplementary C−H functionalization of a pharmacologically relevant compound. Angew. Chem. Int. Ed. 58, 18535–18539 (2019).

44. Naini, A., Sasse, F. & Bronstrup, M. The intriguing chemistry and biology of soraphens. Nat. Prod. Rep. 36, 1394–1411 (2019).

45. Shrestha, Y., Volkath, S. C., Elich, 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 A1: 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 precursors directed biosynthesis. Chem. Comm. 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. Ziekle, R., Ligon, J. M. & Molnár, I. Heterologous production of the antifungal 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. Kocher, O. et al. Mikrobiologisches Verfahren zur Herstellung agrarchemisch verwendbarer mikrobieller makrozyklischer Lactonederivate. EP 358606 A2 (1990).
**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.