ALTEX preprint
published July 2, 2019
doi:10.14573/altex.1904031

Research Article

Machine Learning Prediction of Cyanobacterial Toxin (Microcystin) Toxicodynamics in Humans

Stefan Altaner[1], Sabrina Jaeger[2*], Regina Fotler[1], Ivan Zemskov[3], Valentin Wittmann[3], Falk Schreiber[2,4] and Daniel R. Dietrich[1]*

[1]Human and Environmental Toxicology, University of Konstanz, Konstanz, Germany; [2]Life Science Informatics, University of Konstanz, Germany; [3]Organic and Bioorganic Chemistry, University of Konstanz, Germany; [4]Faculty of IT, Monash University, Melbourne, Australia

Abstract

Microcystins (MC) represent a family of cyclic peptides with approx. 250 congeners presumed harmful to human health due to their ability to inhibit ser/thr-proteinphosphatases (PPP), albeit all hazard and risk assessments (RA) are based on data of one MC-congener (MC-LR) only. MC congener structural diversity is a challenge for the risk assessment of these toxins, especially as several different PPPs have to be included in the RA. Consequently, the inhibition of PPP1, PPP2A and PPP5 was determined with 18 structurally different MC and demonstrated MC congener dependent inhibition activity and a lower susceptibility of PPP5 to inhibition than PPP1 and PPP2A. The latter data were employed to train a machine learning algorithm that should allow prediction of PPP inhibition (toxicity) based on MCs 2D chemical structure. IC_{50} values were classified in toxicity classes and three machine learning models were used to predict the toxicity class, resulting in 80-90% correct predictions.

1 Introduction

Harmful (toxic) cyanobacterial blooms have become an important concern with regard to drinking water quality and safety. Some prominent examples of the latter are a bloom that affected almost 1000 km of the Barwon-Darling River, New South Wales, Australia, in November and December 1991 (Bowling and Baker, 1996), the deaths of renal dialysis patients in 1996 in Caruaru, Brazil (Azevedo et al., 2002), or the most recent closing of the drinking water supply for the inhabitants of Toledo, Ohio, USA, resulting from recurrent *Microcystis aeruginosa* blooms in Lake Erie (Berry et al., 2017). Of concern is the impression that cyanobacterial blooms in surface waters appear to be increasing with climate change (Huisman et al., 2018). Of importance in conjunction with toxic cyanobacterial blooms is that several different toxins and congeners of a given toxin (e.g. microcystins (MC)) can co-occur in a given bloom (Dietrich and Hoeger, 2005), toxin concentrations per cyanobacterial cell can change > ten-fold within a short time span of a bloom event (Wood et al., 2011), and that increased temperature may provide a growth advantage for toxin producing species (Kleinteich et al., 2012).

Microcystins, produced by several cyanobacteria species e.g. *Microcystis spp*., *Dolichospermum spp.* or *Planktothrix spp.* in water bodies worldwide (Preece et al., 2017), appear to be one of the toxins most frequently associated with drinking water, food supplement and/or food contamination and have resulted in human health morbidities and mortalities. Structurally, MC are cyclic heptapeptides consisting of common L-amino acids, but also uncommon and unique amino acids. Their general structure is cyclo([D-Ala1]-[L-X2]-[β-D-MeAsp3]-[L-Z4]-[Adda5]-[γ-D-Glu6]-[Mdha7]). X and Z stand for variable L-amino acids, while β-D-MeAsp is *erythro*- β -D-methylaspartate, ADDA is (2*S*,3*S*,8*S*,9*S*,4*E*,6*E*)-3-amino-9-methoxy-2,6,8-trimethyl-10-phenyl-4,6-decadienoic acid and Mdha is *N*-methyldehydroalanine. The variable positions, along with various (de)methylation sites (Fig. 1[1]), provide for currently 248 known MC congeners (Spoof and Catherine, 2017), albeit new MC congeners are continuously discovered. However, contrary to recent stipulations (Huisman et al., 2018), the toxicity is known but for a very few of the 248 MC congeners.

Indeed, the World Health Organization (WHO) provisional guideline value of 1 µg/L for the risk assessment of MC in drinking water (WHO, 2017), is based entirely on the toxicological data of MC-LR and the assumption that MC-LR is the most toxic of the MC congeners known. This WHO guideline value heavily relies on a 90-day toxicity study in mice (Fawell et al., 1999). Toxicity can be split into two critical components: toxicokinetics (cellular uptake, distribution, metabolism and elimination) and toxicodynamics (the interaction with cellular molecules resulting in an observed adverse outcome)

* authors contributed equally

Received April 3, 2019; Accepted June 27, 2019;
Epub July 2, 2019; © The Authors, 2019.

ALTEX 36(#), ###-###. doi:10.14573/altex.1904031

Correspondence: Daniel R. Dietrich, PhD
Faculty of Biology, University of Konstanz,
Universitätsstraße 10, 78457 Konstanz, Germany
(daniel.dietrich@uni-konstanz.de)

[1] doi:10.14573/altex.1904031s

ALTEX preprint
published July 2, 2019
doi:10.14573/altex.1904031

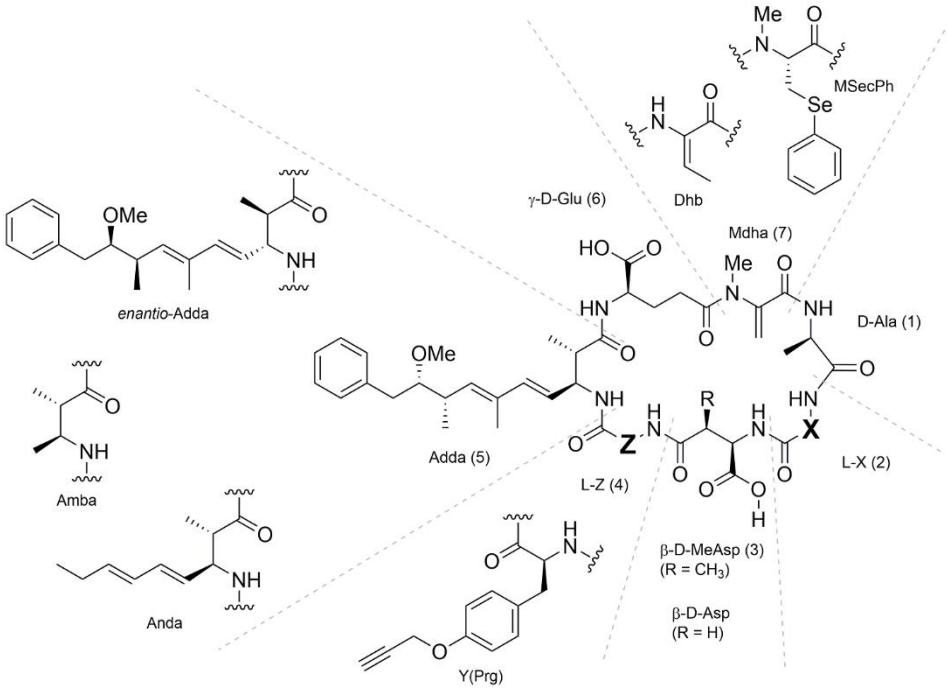

Fig. 1: Consensus structure of microcystins and the synthetic variations produced for this study
The dashed lines represent the single amino acids of the heptapeptide structure. Further details can be found in Table S1. Amba = (2*S*,3*S*)-3-amino-2-methylbutanoic acid, Anda = (2*S*,3*S*,4*E*,6*E*)-3-amino-2-methylnona-4,6-dienoic acid, Dhb = (*E*)-2-amino-2-butenoic acid, MSecPh = *N*-methyl-*Se*-phenyl-L-selenocysteine, Prg = propargyl

(Dellafiora et al., 2018; EFSA PPR Panel et al., 2018). Cellular uptake of MC is primarily governed by organic anion polypeptide (OATP) and thus by the type and level of OATP expression in a given cell, as well as by the affinity and capacity of the respective OATP for transporting the different MC congeners (Fischer et al., 2010). Cellular export of MC (conjugates) is still under debate, as involved exporters have so far not been unambiguously determined. However, the comparison of rodents (mouse and rat) with humans demonstrated that rodents are poor surrogates for humans specifically with regard to type of OATP expressed in the various tissues and the affinity and capacity of expressed OATPs for specific MC congener transport (Feurstein et al., 2011). The fact that humans demonstrated major differences in OATP expression and thus susceptibility to MC (Fischer et al., 2010) only compounded the fact that current risk assessment premises, based on surrogate species and one single MC congener e.g. the WHO guideline value, could severely underestimate the potential toxicities of MC due to their congener specific kinetics.

Similarly, questions arose with regard to the toxicodynamics of MC congeners. Indeed, MC are very potent inhibitors of the catalytic subunits of ser/thr-protein phosphatases (PPP), albeit MC congeners differ with regard to their PPP inhibition capability (Hastie et al., 2005; Mackintosh et al., 1990; Hoeger et al., 2007). The family of PPPs in humans has seven members (PPP1, PPP2A, PPP2B (Calcineurin), PPP4, PPP5, PPP6 and PPP7), whose catalytic subunits are structurally similar (Shi, 2009), display protein sequence homology of up to 65% (checked with Clustal Omega), and have defined substrate specificities and therefore various functions (Pereira et al., 2011). Most of the PPPs are expressed ubiquitously albeit at different levels in the various organs, while PPP7 is specific to retina and brain (Cohen, 2004). Inhibition of PPPs by MC occur *prima facie* via reversible followed by covalent binding of MC to the catalytic subunit of the respective PPP (Mackintosh et al., 1990).

Dysregulated phospho-protein homeostasis, subsequent to PPP inhibition, thus hyperphosphorylation of numerous phosphate-regulated enzymes and the deregulation of fundamental cellular processes, e.g. disruption of the cytoskeleton, represents the toxicodynamic process. As PPP differ in their susceptibility to inhibition by MC (Hoeger et al., 2007) and congeners differ in their capacity to inhibit specific PPP (Hoeger et al., 2007), the observed toxicity as manifested in the respective organs (liver, brain, kidney) are not only the result of MC toxicokinetics but also of toxicodynamics (Fischer et al., 2010). However, to date research on MC toxicodynamics primarily focused on the interaction of MC congeners with PPP1 and PPP2A, whereby only a select few MC congeners (predominantly MC-LR, -RR, -LA, and -LF) were tested (Hoeger et al., 2007; Garibo et al., 2014).

In view of the ever increasing number of MC congeners identified (Spoof and Catherine, 2017), yet lacking availability to synthetize these in sufficient purity and amount for *in vitro*, nor *in vivo*, testing, an *in silico* approach using toxicodynamic data could provide for a first step towards a better "toxicity assessment with relevance for humans" of uncharacterized MCs. However, as there are yet insufficiently robust testing systems to address the toxicokinetic component of toxicity, only the toxicodynamic component, i.e. PPP inhibition, was addressed in the work presented. Indeed, although a number of *in vitro* models have been put forward that allow studying the uptake of MC congeners (Fischer et al., 2010; Monks et al., 2007; Feurstein et al., 2009, 2010, 2011), studying the efflux from cells is much more difficult as intracellular MC would kill the cells before a proper efflux model could be established. Thus, despite recent advances in studying MC efflux using in vitro membrane vesicle approaches (Kaur et al., 2019), complete kinetic models encompassing influx and

efflux kinetics of MCs have yet been impossible to establish. Accordingly, the aim of this study was to develop a comprehensive dataset of toxicodynamics i.e. the PPP inhibitory capacities of a limited number of MC congeners. These *in vitro* data were then used as a comparative basis driving an *in silico* approach using machine learning (ML). Thus, the PPP inhibition capacity (toxicity) of 18 structurally diverse MC congeners was determined using ser/thr-PPP (PPP1, PPP2A and PPP5) in a colorimetric protein phosphatase inhibition assay. For the latter, a number of synthetic MC derivatives were generated according to previously published procedures (Zemskov et al., 2017; Fontanillo et al., 2016). Among these was a variant with modified stereochemistry at the Adda5 residue (i.e., the enantiomer of Adda was used) ([*enantio*-Adda5]-MC-LF) as well as variants with simplified residues at the Adda5 position ([Anda5]-MC-LY(Prg) and [Amba5]-MC-LY(Prg)). The modified amino acids in these synthetic MCs in positions other than X2 and Z4 are indicated by adding the modification in square brackets before the name of the MC-XZ derivative. MC-LY(Prg) denotes variants with L-leucine in position X2 and *O*-propargylated L-tyrosine in position Z4 (Fig. **1** and Table S1[1]). Results were classified into three categories (toxic, less toxic, non-toxic) and the toxicity predicted based on chemical structure via the ML approach described below.

Machine learning is widely used in the field of bioinformatics to predict bioactivity or molecular properties (e.g. solubility) of compounds, protein folding, etc. Despite the latter advances, it is still difficult to employ ML in pharmacology or toxicology, as datasets are often smaller and more heterogeneous compared to datasets from other domains (Wu et al., 2018). Indeed, although ML has been employed for the prediction of cyanobacterial blooms based on satellite data (Chang et al., 2014) or the production of toxins based on environmental factors (Taranu et al., 2017), ML has so far not been used to predict the toxicity of MC congeners. To encode molecules or proteins for ML, a fixed size numerical vector is needed (Wu et al., 2018). One approach to encode molecules and proteins is Mol2vec (Jaeger et al., 2018) and ProtVec (Asgari and Mofrad, 2015), respectively, which are inspired by natural language processing. Both approaches are based on the Word2vec approach, which is generating vector representation of words to capture semantic meaning (Mikolov et al., 2013). The vectors are obtained by training a deep neural network based on a database of different text (so-called corpus) and results in a dense, high-dimensional representation of words. This procedure is a pre-training, resulting in a look-up table of words and vectors, which can be extracted later for various applications (e.g. ML). To encode molecules with the Mol2vec approach a large corpus of a database or collection of molecular structures has to be generated (Figure 2A). The Morgan fingerprint is calculated for each molecule, but instead of hashing identifiers of substructures in a bit vector, identifiers (or words) are extracted and ordered to a sentence to represent a molecule. By this procedure, a molecular lookup table of molecular substructures and vectors is generated, which is able to capture chemical relationship between substructures. To represent a new molecule, a Morgan fingerprint is generated, identifiers are extracted, and looked up in the pre-trained model. Then, all substructure vectors are summed up to represent one molecule (Figure 3), which results in a fixed size numerical representation of the molecule (Jaeger et al., 2018).

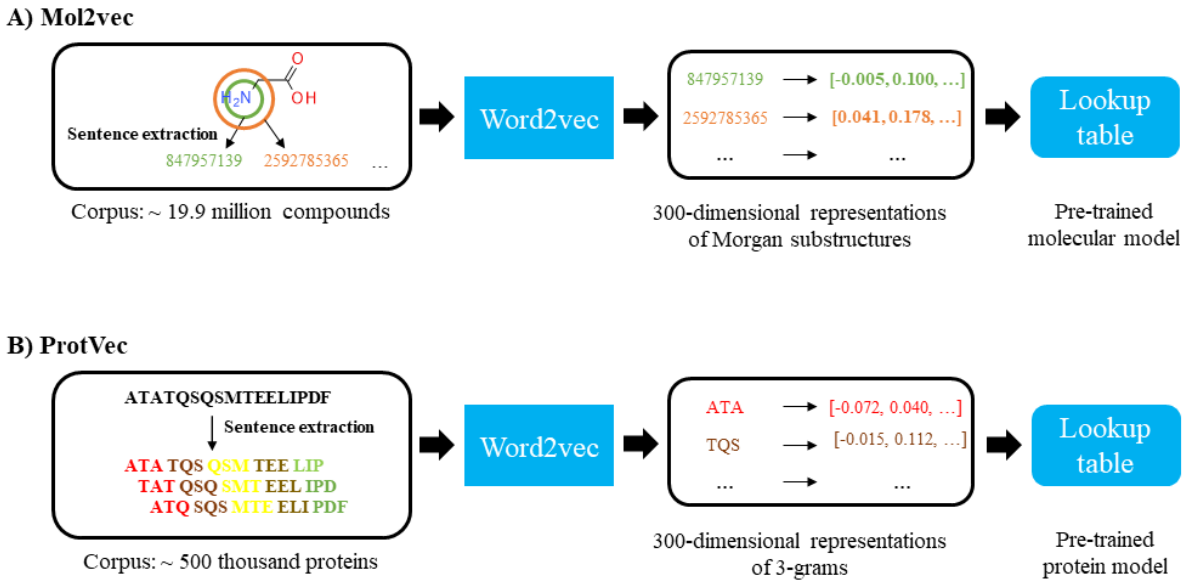

Fig. 2: Application of Word2vec to molecules and proteins
The procedure results in a lookup table, where 300-dimensional vectors can be extracted to have a numerical representation. **A)** For molecules, identifiers generated by Morgan fingerprint are considered as words, and their ordering as a sentence or molecule. **B)** For proteins, all possible 3-grams are considered as words, and by applying a sliding window over protein sequence, three sentences are generated to represent a protein. Modified from Jaeger et al. (2018).

To encode a protein with the ProtVec approach (Figure 2B), a large corpus of a database or collection of protein sequences has to be generated. Therefore, all possible n-grams of a protein sequence are generated by applying a sliding window over a protein sequence. This results in n-sentences to represent a protein. Those n-sentences are then used to generate the lookup table of protein n-grams. To represent a new protein, all n-grams are generated, looked up in the pre-trained model, and summed up to represent a new protein (Figure 3), which results in a fixed size numerical representation of the protein (Jaeger et al., 2018; Asgari and Mofrad, 2015).

ALTEX preprint
published July 2, 2019
doi:10.14573/altex.1904031

PPP inhibition data (IC_{50}'s) gained from the *in vitro* assays for the different MC congeners (Table 1) were then classified into three classes of "toxicity" (Table 2, Figure 3) and ML models generated, using the encoded MC congeners and PPP vectors (Figure 2 and 3). These were then trained with different features and classifiers to classify MC congeners into the toxicity classes, as shown in the ML flowchart (Figure 4). The latter approach thus allowed to predict the toxicity of MC congeners and to compare the predictions with the true findings from the *in vitro* PPP inhibition assays.

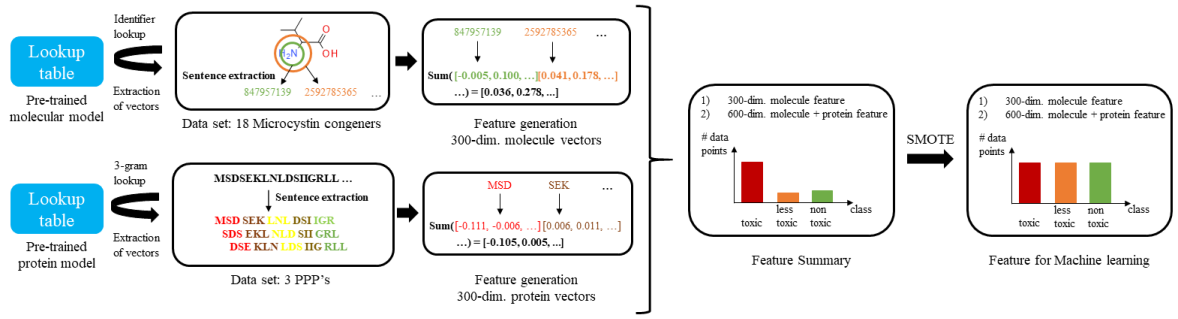

Fig. 3: Workflow for feature generation for toxicity classification
After downloading a pre-trained model for molecular structures and pre-training a protein model, vectors are extracted for substructures of 18 Microcystin congeners and the 3-grams of the 3 PPPs. To represent a molecule or protein, the respective vectors are summed up. Then, either 300-dimensional molecule vector is extracted, or combined with protein vector, to build a 600-dimensional vector. As the data set is highly imbalanced for the different classes, synthetic minority oversampling technique (SMOTE) is applied, to have the same number of compounds for each class. Modified from Jaeger et al. (2018).

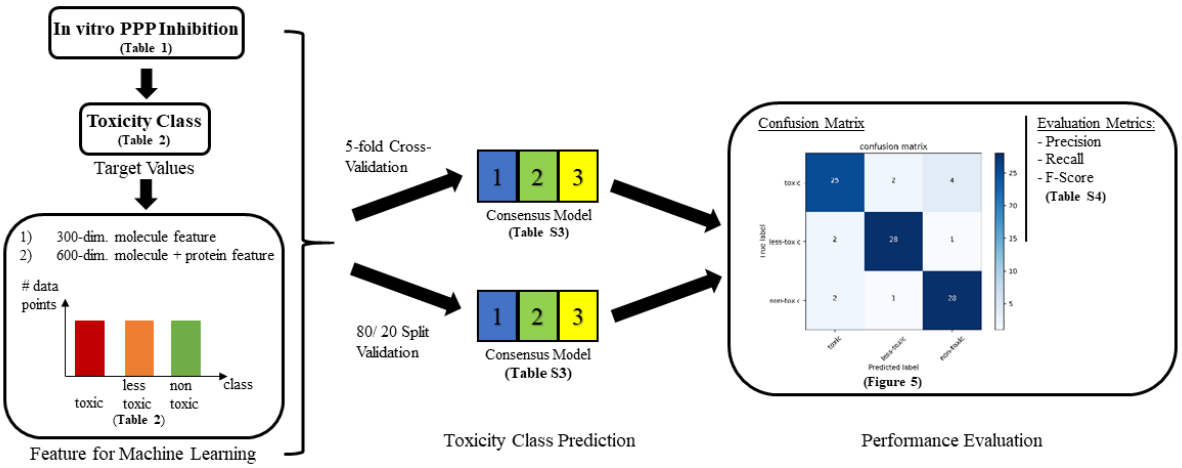

Fig. 4: After feature generation and pre-processing of the data, respective target values (toxicity class) are combined with the feature vector and used for a machine learning classification
Two different validation methods were used. For both validation methods, three machine learning models were set up, and majority voting was used for final prediction, building a so-called consensus model. Afterwards, performance was evaluated with a confusion matrix and evaluation metrics.

2 Materials and Methods

2.1 Materials

Microcystins were obtained either from Enzo Life Sciences (MC-RR, -LR, -YR, -WR, -LA, -LY, -LF, -LW, -HilR, -HtyR, [β-D-Asp3]-MC-RR and [β-D-Asp3]-MC-LR) or were synthesized (MC-LY(Prg), [*enantio*-Adda5]-MC-LF, [Anda5]-MC-LY(Prg), [Amba5]-MC-LY(Prg), [MSecPh7]-MC-LY(Prg)) using previously published procedures (Zemskov et al., 2017; Fontanillo et al., 2016). [β-D-Asp3, Dhb7]-MC-RR was a gift from Judith Blom (University of Zürich, Switzerland).

Microcystins were dissolved in pure methanol to 100 µM. Actual concentrations were determined using UV spectroscopy at 238 nm (using the extinction coefficient of MC-LR of 39800 mol l^{-1} cm^{-1}) and stocks were stored at -20°C until used for serial dilutions. Stocks of [Anda5]-MC-LY(Prg), [Amba5]-MC-LY(Prg) were quantified by dissolving weighed amounts in an appropriate volume of methanol, as photometric quantification was not possible due to lack of the characteristic absorption peak at 238 nm (missing ADDA chain).

rPPP1 (rabbit skeletal muscle) was obtained from New England Biolabs (P0754L, product discontinued). hPPP2A (human red blood cells) was from Promega (V6311, product discontinued). pET32a(+)-TrxA-6His-hPPP5 was generated by GenScript using human PPP5 sequence (NCBI Accession NP_006238.1) with GenScripts codon optimization for *E.coli*.

2.2 Expression of 6xHis-hPPP5 in *E. coli*

pET32a(+)-TrxA-6His-hPPP5 was transformed into chemical competent BL21-CodonPlus(DE3)-RP *E.coli* cells (Agilent, 230255) via heat-shock. After selection on ampicillin/chloramphenicol-agar, a single colony was picked and cultivated in 2 ml LB-ampicillin/chloramphenicol (amp/cam) medium for 6h. Afterwards, the pre-culture was added to 500 ml of TB-amp/cam (Terrific Broth supplemented with ampicillin and chloramphenicol) medium and incubated over night at 37°C while shaking (220 rpm). The following day OD_{600} was measured to ensure growth in the exponential phase (OD_{600} = ~4). Controls were performed using heat-shock treated BL21-CodonPlus(DE3)-RP *E. coli* cells without plasmid.

2.3 Purification of 6xHis-hPPP5

Cultures of transformed BL21-CodonPlus(DE3)-RP *E. coli* cells were centrifuged at 5000×g for 10 min at 4°C. Pellets were washed with 10 ml STE buffer (10 mM Tris-HCl pH 8.0, 100 mM NaCl, 1 mM EDTA) prior to resuspension in 10 ml cold PBS buffer + 1% protease inhibitor cocktail (Sigma, P8849). Cell lysis was achieved using a Branson Sonifier 250 with five times 20 pulses. Samples were cooled on ice between each set of 20 pulses. Lysate was cleared from cell debris by centrifugation for 45 min at 18000×g at 4°C. The cleared lysate was incubated with 2.5 ml (50:50 slurry) equilibrated Ni-NTA-agarose beads (Biozym, 2631105) on an overhead-tumbler at 4°C over-night. After centrifugation at 4°C and 800×g for 5 min, the supernatant was discarded and beads were washed three times with 2.5 ml wash buffer (50 mM NaH_2PO_4, 300 mM NaCl, 60 mM imidazole). Elution was performed by incubating the beads for 10 min at 4°C on an overhead-tumbler in 2 ml elution buffer (50 mM NaH_2PO_4, 300 mM NaCl, 100 mM imidazole). Samples were taken from each step for subsequent SDS-PAGE analysis. To avoid high imidazole concentrations in the final sample, buffer exchange via dialysis against storage buffer (20 mM TrisHCl pH 8, 100 mM NaCl) was performed using 2 L volumes for about 18 h with two buffer exchanges. 10% glycerol was added to the final samples before aliquoting and liquid nitrogen snap-freezing. All steps were either performed on ice or within a cooling room (4°C). Samples were stored at -80°C until further use. In addition, one analogous purification was performed using the heat-shock treated BL21-CodonPlus(DE3)-RP *E.coli* without plasmid ("empty expression") for control purposes.

2.4 SDS-PAGE analysis

SDS-PAGE analysis was performed using 10% SDS-gels in a Bio-Rad Mini-PROTEAN vertical electrophoresis system. 6 µl 6 × SDS-sample buffer was added to 30 µl sample and the samples loaded into the gel pockets. Samples were not corrected for protein amount, but for the volume of the original fraction in such a fashion that consistently 0.3% of the original fractions were used (filled up to 30 µl with MilliQ). Gels were run at 100 V for 60 min and stained using colloidal Coomassie solution at 4°C while shaking over-night. Images were taken using a desktop scanner.

2.5 Mass spectrometry

Mass-spectrometry was employed to confirm the identity of the TrxA-6×His-hPPP5 expressed. Samples were run on SDS-gels as described above and bands considered to contain TrxA-6×xHis-hPPP5 were cut out, divided into ~1 mm² squares and submitted to the Proteomics Core Facility of the University Konstanz. Proteomic analyses of trypsin digested fragments were carried out with an LTQ Orbitrap Discovery (Thermo Fisher Scientific, Bremen Germany) coupled to an Eksigent 2D-nano HPLC (Eksigent, USA). Data were analyzed using Mascot software (Matrix Science).

2.6 Phosphatase activity assay

The phosphatase activity of the TrxA-6xHis-hPPP5 fraction (hPPP5) was assayed using a colorimetric phosphatase activity assay with *para*-nitrophenylphosphate (pNPP) as substrate and buffers according to Heresztyn and Nicholson (Heresztyn and Nicholson, 2001). Purified hPPP5 was tested undiluted and in 1:1 serial dilutions (total 11 concentrations, highest dilution 1:1024) using the enzyme diluent buffer (52 mM Tris pH 7, 2 mM $MnCl_2$, 0.5 mg/ml BSA, 1 mM DTT, 0.5 mM NaOAc, 123.5 µM EGTA) for dilutions. 20 µl of each dilution was pipetted into a well of a polystyrene flat bottom 96-well plate to which 200 µl of the substrate solution (62.5 mM Tris pH 8.1, 26 mM $MgCl_2$, 0.2 mM $MnCl_2$, 0.5 mg/ml BSA, 2 mM DTT, 1 mM NaOAc, 24 mM pNPP) was given. Testing was carried out in technical triplicates. Color development at 37°C and 405 nm was measured every 10 minutes over a period of 4 h. Linear regressions of each dilution over time were plotted and slopes obtained. Slopes obtained were compared to corresponding slopes of rPPP1 (protein amount and specific activity). Accordingly, this allowed calculation of the volume of the hPPP5 fraction corresponding to three U of rPPP1 (hPPP5: 0.83 U/µl, hPPP1: 2.5 U/µl). U here is defined as the amount of enzyme needed to dephosphorylate 1 nmol of pNPP in 1 minute at 30°C.

2.7 Colorimetric protein phosphatase inhibition assay (cPPIA)

The assay employed is based on the previously published procedures (Heresztyn and Nicholson, 2001; Fischer et al., 2010). Serial dilutions of each MC congener were produced in MilliQ in LC-vials, whereby the most concentrated MC sample contained a maximum of 5% methanol. 20 µl of each dilution was pipetted into a polystyrene flat-bottom 96-well plate in triplicates. Three U of each phosphatase were diluted in 2120 µl enzyme dilution buffer and 20 µl of the solution were added to each well (corresponding to about 0.07 U per well). Enzyme solution (20 µl) in addition to 20 µl of MilliQ (three replicates) served as 100% activity control (no inhibition) while 20 µl enzyme dilution buffer lacking enzyme (three replicates) as well as 20 µl MilliQ served as background control (no enzyme activity). The plate was incubated at 37°C for 5 min to ensure interaction of microcystins with the PPP. To each well 200 µl of substrate solution was pipetted and the plate was immediately read at 405 nm using an Infinite 200 Pro microplate reader (Tecan, Männedorf, Switzerland). The plate was then incubated at 37°C for 3h before being measured again at 405 nm. PPP activity was calculated by subtracting the start value (0h) from the end value (3h) and compared to 100% activity. IC_{50} were calculated using GraphPad Prism 5.0 software via a 5-PL non-linear regression with anchorage points and constraints between 100% and 0%. Replication: n = at least 3 for

PPP2A and at least 5 for PPP1 and PPP5, each in technical duplicates or triplicates. The analyses of [Amba5]-MC-LF and [Anda5]-MC-LF had 5 biological replicates but no technical triplicates due to shortage of pure testing material.

2.8 Data analyses, pre-processing, statistics and machine learning

Data analyses were carried out using Microsoft Excel Professional Plus 2013, while GraphPad Prism 5 was used for statistics and data representation. Data pre-processing and machine learning were carried out using python version 3.6.6.

2.9 Machine learning (ML)

The PPP inhibition capabilities (toxicodynamics) of the different MC congeners, expressed as IC_{50}'s (Table 1), were used to train a ML model (Figures 2-4). In order to allow for the adaption into the ML model, the two primary factors, i.e. MC congeners (molecules) and PPP (proteins), had to be transformed to numerical vectors (Figure 2). To transform molecules to a vector, the Mol2vec approach, as described in Jaeger et al. (2018) was used with a pre-trained model[2]. To transform proteins to a vector, the ProtVec approach (Asgari and Mofrad, 2015) was used (Figure 2) and a model trained on UniProt sequences as described in (Jaeger et al., 2018). Subsequently, the models were applied to encode MC congeners and the PPPs (UniProt ID: PPP1 (P62136), PPP2A (P67775) and PPP5 (P53041)) as vectors.

PPP inhibition data (IC_{50}'s) of the different MC congeners (Table 1) were classified into three classes of "toxicity" (Table 2, Figure 3 and 4). The original data set consisted of 47 data points of which 31 data points were classified as toxic, seven were classified as less-toxic and nine were classified as non-toxic. This classification chosen was arbitrary, albeit the most toxic classification includes MC congeners with relevance to human intoxications and the WHO guideline (Berry et al., 2017; Dietrich and Hoeger, 2005; Azevedo et al., 2002; WHO, 1999). As this is a rather small data set for ML and sufficient samples are crucial for ML performance, Synthetic Minority Oversampling Technique (SMOTE) was applied to mimic a balanced dataset and thereby increase prediction performance of minority classes (Chawla et al., 2002). Indeed, SMOTE generates new, artificial data points for minority class by variation of the feature vector representing original data points. SMOTE implementation in imbalanced learn (version 0.3.3) (Lemaître et al., 2017) was used with standard settings and a ratio of 1.0. This procedure resulted in 31 data points per class, resulting in 93 data points in total (Table 2).

Three different ML models were trained with different features and classifiers to classify MC congeners into the toxicity classes. Two models were trained with a random forest (RF) classifier implemented in scikit-learn (version 0.8.0) (Pedregosa et al., 2011) and the XGBoost implementation (version 0.8.0) of Gradient Boosting Machines (GBM) classifier (Friedman, 2001), respectively. As feature, vectorized structural data of the congeners (Mol2vec) was used. Additionally, one model was trained with a random forest classifier with molecular and protein data as feature (Table S3[1], Figures 3 and 4). To merge molecule and protein information into one vector, structural data of molecules was vectorized with Mol2vec and protein data was vectorized with ProtVec and then concatenated. Subsequently, hyper-parameters were tuned (Table S3[1]) to derive the best model. The final model used majority voting of these three models (Mol2vec with RF, Mol2vec with GBM and Mol2vec + ProtVec with RF) to classify a compound (Figure 4).

For training and evaluation of an ML algorithm, the data set had to be split into a training and a test set. Those two sets had to be strictly separated, because when data points from the training set were used in the test set, the evaluation would always result in high performance, because the model already knew the data point. Here, two procedures were applied to split the data set: 1) using 80% of the data points (75) for training and 20% of the data points (18) for testing the performance and 2) by five-fold cross validation (Figure 4). Applying five-fold cross validation results in four-folds of the data set used for training and one-fold of the data set used for testing. This procedure is repeated, until every fold was once the test set, resulting in five ML models. This procedure has the advantage, that a standard deviation of performance between the models can be calculated to get a better estimation of model performance and robustness (Table S4). For both procedures data points are randomly assigned to the respective set or fold. For this reason, each procedure and ML was repeated 50 times, to test whether performance was robust and independent of the random assignment of data points for training.

To finally estimate performance of the ML model, different performance metrics were employed (precision, recall, F-score, see Table S4 and Figure S1[1]). In addition, the confusion matrix (Figure 5) was checked, to estimate how many and which molecules were classified correctly or falsely. Performance metrics and confusion matrix were used as implemented in scikit-learn (Pedregosa et al., 2011).

3 Results

Full-length human PPP5 was bacterially expressed in BL21 Codon Plus *E. coli* with several attached tags: Thioredoxin A (TrxA) for solubility, 6-Histag for purification, S-Tag for a possible second purification and a thrombin-site (TrxA-6His-S-PPP5). PPP5 identity was confirmed using mass spectrometry after purification using Ni-NTA beads (Table S2[1]). To ensure that the observed activity was due to PPP5, bacteria without a plasmid were grown, purified and tested. These purified proteins did not show PPP activity (Figure S2[1]).

To develop a dataset of MC congener dependent "toxicity" (PPP inhibition activity) for the ML model, 18 different MC congeners were tested in three PPP (rPPP1, hPPP2A, hPPP5 expression #1). The 18 MC congeners spanned the known spectrum of hydrophobicity, had different molecular weights, and contained common as well as unusual modifications of the consensus structure (Figure 1 and Table S1[1]). The *in vitro* PPP inhibition assays provided for well fitted (R^2) concentration-inhibition response curves (Table 1, Figures S3-5[1]), whereby the derived IC_{50}'s were subsequently used for the ML approach (Figures 2-4).

[2] Mol2vec - an unsupervised machine learning approach to learn vector representations of molecular substructures. https://github.com/samoturk/mol2vec (accessed 08.08.2018)

Tab. 1: IC_{50}s of the tested MC congeners on rPPP1, hPPP2A and hPPP5. IC_{50} are calculated with GraphPad Prism 5 after 5PL-nonlinear regression of at least three (hPPP2A) or five (rPPP1 and hPPP5) individual replicates using technical duplicates or triplicates. n.d. not determined (PPP2A not available any more, discontinued by manufacturer).

Congener	rPPP1			hPPP2A			hPPP5		
	IC_{50} (nM)	CI_{95} (nM)	R^2	IC_{50} (nM)	CI_{95} (nM)	R^2	IC_{50} (nM)	CI_{95} (nM)	R^2
MC-RR	1.5	1.3 – 1.8	0.95	1.6	1.4 – 1.7	0.99	11.7	8.3 – 16.5	0.96
MC-LR	0.3	0.2 – 0.4	0.93	0.5	0.4 – 0.5	0.99	5.1	4.0 – 6.6	0.97
MC-YR	1.3	1.2 – 1.5	0.99	n.d.	n.d.	n.d.	5.1	4.3 – 6.1	0.99
MC-WR	1.2	1.0 – 1.5	0.94	1.0	0.8 – 1.1	0.97	5.6	4.2 – 7.6	0.97
MC-LA	1.9	1.4 – 2.7	0.86	0.7	0.5 – 0-9	0.93	6.1	4.3 – 8.7	0.96
MC-LY	0.8	0.7 – 0.9	0.99	n.d.	n.d.	n.d.	4.1	3.1 – 5.4	0.97
MC-LF	2.0	1.5 – 2.6	0.90	1.4	1.3 – 1.4	0.99	4.7	3.5 – 6.3	0.97
MC-LW	1.2	1.0 – 1.4	0.97	0.7	0.7 – 0.8	0.99	2.5	2.0 – 3.2	0.98
MC-HilR	0.6	0.5 – 0.8	0.99	n.d.	n.d.	n.d.	4.2	3.5 – 5.1	0.99
MC-HtyR	0.7	0.6 – 0.8	0.99	n.d.	n.d.	n.d.	4.7	3.6 – 6.0	0.96
[β-D-Asp3]-MC-RR	45.0	39.3 – 51.6	0.99	n.d.	n.d.	n.d.	167.1	131.8 – 211.8	0.97
[β -D-Asp3]-MC-LR	0.9	0.7 – 1.0	0.99	n.d.	n.d.	n.d.	10.2	8.3 – 12.5	0.99
[β -D-Asp3, Dhb7]-MC-RR	62.0	51.7 – 74.3	0.96	84.3	80.7 - 87.8	0.99	877.1	692.6 – 1111	0.97
MC-LY(Prg)	1.7	1.3 – 2.2	0.95	0.4	0.2 – 0.3	0.99	1.7	1.2 – 2.6	0.95
[MSecPh7]-MC-LY(Prg)	1.9	1.6 – 2.4	0.97	0.9	0.7 – 1.1	0.94	18.2	10.7 – 31.1	0.91
[*enantio*-Adda5]-MC-LF	-	-	-	-	-	-	-	-	-
[Amba5]-MC-LY(Prg)	520817	449800 – 603048	0.98	2135	1991 - 2291	0.99	54063	37431 - 78087	0.95
[Anda5]-MC-LY(Prg)	1724	1434 - 2072	0.98	n.d.	n.d.	n.d.	2420	1690 - 3467	0.96

The comparison of IC_{50}'s obtained with MC congeners in the three PPPs tested, demonstrated that of the ten MC congeners available for testing in PPP1 and 2A, five MC congeners had a comparable IC_{50} values, while five MC congeners were more selective towards PPP2A, possibly suggesting a slightly higher sensitivity of PPP2A toward MC congeners (Table 1 and Figure S5[1]). In contrast, several-fold higher concentrations of MC congeners were necessary to achieve 50% inhibition of PPP5 phosphatase activity. Notable exceptions to the latter were MC-LY(Prg) and [Amba5]-MC-LY(Prg), to which the PPP5 susceptibility to inhibition was comparable and but lower than observed for PPP1.

The importance of structural differences with regard to binding to the catalytic subunit of PPPs was dramatically demonstrated with the comparison of MC-LF and the *de novo* synthetized [*enantio*-Adda5]-MC-LF. While MC-LF inhibited all three PPPs tested, the corresponding [*enantio*-Adda5]-MC-LF had no PPP inhibitive activity at all (Table 1). In contrast, other structurally similar derivatives, i.e. MC-LY(Prg) and [MSecPh7]-MC-LY(Prg), both having a propargyloxy residue at position 4 (Figure 1) of the phenylalanine moiety, show only slightly reduced PPP inhibiting activity, if any, when compared to the parent MC-LF. However, if the Adda-residue is shortened to [Amba5]-MC-LY(Prg) or [Anda5]-MC-LY(Prg) (Figure 1 and Table S1[1]) a marked reduction in PPP inhibiting capacity is found (Table 1). The latter observations suggest that structural changes of the amino acid Adda (enantiomeric configuration or shortened Adda-side chain) prohibited or reduced functional interaction with the catalytic subunits and thus inhibition of the PPPs. On the other hand, structural changes to the phenylalanine moiety at position 4 or the Mdha at position 7 had limited impact on PPP activity. Similarly, exchanging leucine in MC-LR and tyrosine in MC-YR for a homoisoleucine (MC-HilR) and a homotyrosine (MC-HtyR) at position 2 had no significant effect on PPP inhibition capacity (Table 1). However, changing the methylation of β-D-MeAsp at position 3 of MC-RR to demethylated [β-D-Asp3]-MC-RR and [β-D-Asp3,Dhb7]-MC-RR, resulted in a decreased PPP inhibition capacity, thus suggesting that structural changes involving L-amino acid residues at position 3 could have an impact on the inhibition of PPPs. When comparing the impact of structural changes of the Adda side-chain at position 5 with changes of β-D-MeAsp at position 3, it appears that the former had a much more pronounced impact on the binding of MC congeners into the catalytic subunit and thus inhibition of PPPs.

The above IC_{50} values were classified in toxicity classes (Table 2) and used for an ML approach. As the distribution of data points was not similar among classes, oversampling was applied, resulting in 31 data points per class, adding up to a total of 93 (Table 2, Figures 3 and 4). The resulting data was used for ML via two different approaches: using 80% of the data for training and 20% for testing of prediction (80/20) and by 5-fold cross validation (CV) with 50-times repetition (see Figure 4, Table S4[1] and 3.9 for details).

Tab. 2: "Toxicity" classes assigned to MC congeners
Classification is based on their PPP inhibitive capabilities (expressed as IC_{50}) for each of the three PPPs tested.

IC_{50} (nM)	Class	Description	Number data points (original)	Number data points after SMOTE
≤10	0	Toxic	31	31
> 10 ≤ 1000	1	Less-toxic	7	31
> 1000	2	Non-toxic	9	31
		Total	**47**	**93**

Subsequently, all data points were used and majority voting of three ML models was employed to predict the toxicity class of every data point (Table S3[1]). Both approaches of splitting data in training and test set performed well for toxicity class prediction with a precision above 0.8 and a recall and F-score mostly above 0.8 (Table S4[1]). Since cross-validation is more suitable for small data sets (Beleites et al., 2013), we primarily focused on CV results, although 80/20 results are provided in Table S4 [1]as well. CV predictions were compared with the true classification according to classified IC_{50} values (Figure 5A).

25 of the 31 toxicity data points were predicted correctly, while two were wrongly predicted as less toxic and four as non-toxic. Interestingly, the misclassified MC congeners primarily involved MC-LF variants. For example, the stereoisomer of MC-LF ([*enantio*-Adda5]-MC-LF) was classified as toxic in PPP2A, despite that it is not toxic. The latter resulted most likely from the fact that the training vector generation was not trained for chirality of the molecules in question. Indeed, the vector generated for [*enantio*-Adda5]-MC-LF and MC-LF would be identical, albeit the values entered into the ML algorithm would read "non-toxic" and "toxic" and thus result in wrong classifications by ML. To test this, the same approach was chosen, yet while omitting the [*enantio*-Adda5]-MC-LF-variant from the training and test sets of the ML algorithm (Figure 5B). In consequence, both natural MC-LF and propargylated MC-LF variants were classified correctly. Moreover, [β-D-Asp3]-MC-LR (in PPP1) was now moved from originally being wrongly predicted as non-toxic, to be now predicted as less toxic, despite its true affiliation in the "toxic" class. Similarly, the prediction of [β-D-Asp3]-MC-LR (in PPP5) moved from non-toxic to toxic, despite its true affiliation in the less-toxic class. Finally, MC-RR (in PPP5) moved from the correct prediction of "less toxic" to the wrong prediction of "toxic". Overall, the ML model trained without the MC-LF stereoisomer performed better, producing fewer false negative predictions.

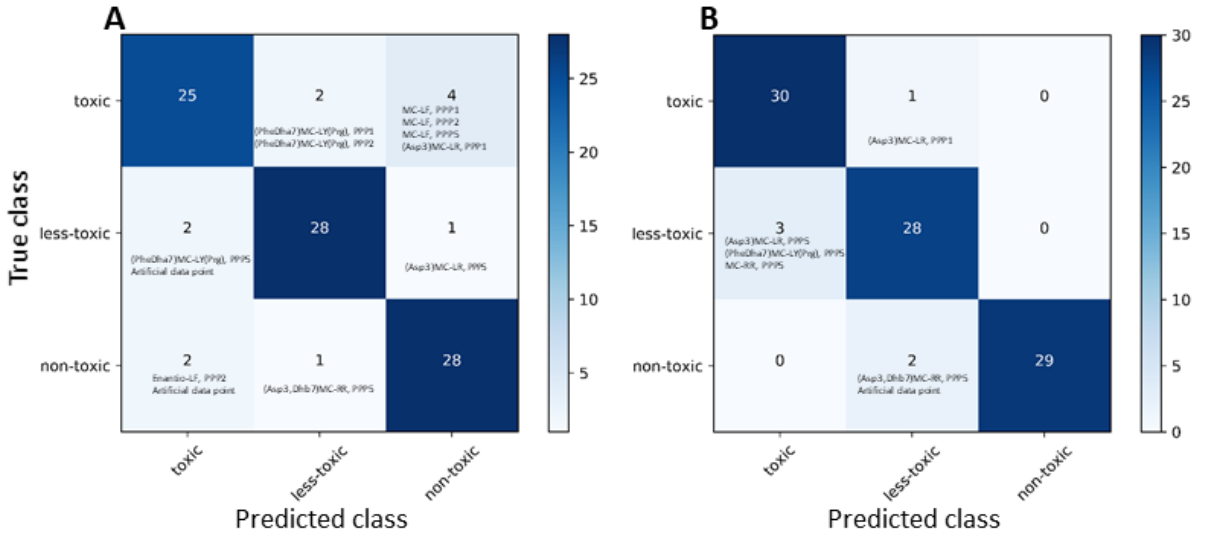

Fig. 5: Confusion matrix of microcystin toxicity prediction using 5-fold cross validation
Either the whole dataset was used for training **(A)**, or the stereoisomer [*enantio*-Adda5]-MC-LF) was dropped for the training of the algorithm **(B)**. The identity of all wrong classifications is given in the individual squares.

4 Discussion

Ser/thr-PPP are an evolutionary very old family of enzymes, with PPP1 and PPP2A being two of the most slowly evolving enzymes (Cohen, 2004). PPP1 usually works together with a regulatory subunit, while the PPP2A-holoenzyme additionally comprises a scaffold subunit (Shi, 2009). In this regard PPP2A is similar to PPP4 and PPP6, whose holoenzymes are also comprised of three subunits, i.e. scaffold, regulatory and catalytic (Brewis et al., 1993; Stefansson et al., 2008). Moreover, PPP4 and PPP6 are more closely related to PPP2A than to any other member of the PPP family (Shi, 2009). Thus, it can be assumed that PPP4 and PPP6 should behave similarly to MC inhibition as PPP2A. PPP5 is the most distant from the other members of the PPP family (Andreeva and Kutuzov, 2001). PPP5 is expressed as a single peptide combining the catalytic domain with a TPR-domain, which interacts with the peptides c-terminus to act as auto-inhibitory domain (Yang et al., 2005).

The catalytic subunits of PPP1, PPP2A and PPP5, representing three members of distinct subfamilies of the PPP family, were used for inhibition assays with various MC congeners. In view of the absence of pure human PPP1 available for *in vitro* testing, PPP1 from rabbit skeletal muscle was employed as it displays 100% protein sequence identity to human PPP1 (hPPP1; analysed with Clustal Omega and Geneious (Biomatters)) and is therefore considered to be equal in structure and enzymatic performance to hPPP1. In contrast, hPPP2A and hPPP5 have only 43% and 37% sequence homology with hPPP1, while hPPP5 and hPPP2A are 40% homologous (analysed with Clustal Omega). Despite sequence differences, the 3D-structures of the respective catalytic subunits align quite well (see Figure S6[1]), suggestive of similar size and structure restrictions for MC interaction with the hydrophobic groove close to the catalytically active center of the respective PPPs.

Indeed, despite that PPPs have been reported to be inhibited by MC earlier (Fischer et al., 2010; Hastie et al., 2005; Mackintosh et al., 1990; Garibo et al., 2014), the data presented here are unique as they compare the MC-mediated inhibition of the catalytic subunits of three different PPPs in parallel with an hitherto unprecedented number of MC congeners, including synthetically derived structural variants. The data demonstrated that MC-RR, -LR, -YR, and [β-D-Asp3, Dhb7]-

MC-RR presented with similar IC_{50} for PPP1 and PPP2A (Table 1), while PPP2A was ≥2-fold more sensitive to the more hydrophobic MC-LW, -LA, -LF and the synthetic MC-LF derivatives (MC-LY(Prg), [MSecPh7]-MC-LY(Prg) and [Amba5]-MC-LY(Prg), see Figure 1). Although ([Amba5]-MC-LY(Prg) displayed an approx. 243-fold higher toxicity towards PPP2A than to PPP1, this was not the case for the other two OPrg-containing congeners ([MSecPh7]-MC-LY(Prg) and [Anda5]-MC-LY(Prg)). Indeed, it has previously been described that MC variants with reduced Adda5-sidechains show a tendency to bind more effectively to PPP2A than to PPP1(Fontanillo et al., 2016).

In principle the MC congener's inhibition capacity of PPP5 followed the same trend as observed for PPP1 and PPP2A, albeit being 4-200-fold lower. With the exception of [*enantio*-Adda5]-MC-LF, showing absence of binding to all three PPPs tested (Table 1), earlier assumptions regarding size and structure restrictions for MC interaction with the respective PPPs could not be corroborated. Indeed, MC congeners apparently do not bind as tightly to the catalytic subunit of PPP5 as to the catalytic subunits of PPP1 and 2A. Exception to the latter, surprisingly, were the synthetic MC-LY(Prg) and [Anda5]-MC-LY(Prg), sharing similar inhibition capabilities in PPP1 and PPP5 (Table 1). Asp3 variants of MC-LR appeared to be of comparable toxicity as MC-LR, while in contrast β-D-Asp3 variants of MC-RR were all dramatically less toxic than MC-RR across all PPPs tested. Although MC-LR is considered to be the most toxic of all congeners (WHO, 2017), this appears to apply only to PPP1 when considering toxicodynamic data. Indeed, MC-LW, MC-LF, MC-LY(Prg) and [MSecPh7]-MC-LY(Prg) presented with comparable inhibiting capabilities as MC-LR in PPP2A. Moreover, in PPP5 MC-WR, -YR, -LY, -LA, -HilR, -HtyR were of comparable toxicity while MC-LF and MC-LY(Prg) were more toxic than MC-LR. The latter suggests that using the toxicity equivalent factors concept (TEF), i.e. all MC congeners equaling in toxicity to MC-LR, would under- and overestimate the potential toxicodynamic capacity present in a given cyanobacterial bloom. Moreover, the fact that MC congeners have been demonstrated to present with significant differences with regard to OATP transport (Fischer et al., 2005; Feurstein et al., 2011), whereby MC-LR and-RR are transported less efficiently than e.g. MC-LA,-LW, or -LF, compounds the problems mentioned with using the TEF as originally proposed by Dietrich and Hoeger (Dietrich and Hoeger, 2005). Indeed, in a realistic setting employing a guidance value of 1 µg $MC\text{-}LR_{equivalent}$ /L for drinking water (Falconer and Humpage, 2005) and using summary detection methods e.g. ELISA (Fischer et al., 2001) without concurrent LC-MS/MS confirmation of MC congeners present (Puddick et al., 2014), could severely under- or overestimate the toxicity of a MC congener mixture in a given water sample contaminated by a toxic cyanobacterial bloom. Indeed, there are several reports of multiple co-occurring MC congeners in a given cyanobacterial bloom (Falconer et al., 1994; Kleinteich et al., 2018; Puddick et al., 2014), thus demonstrating the reality of having to deal with mixture exposures of different toxicities in a human hazard and risk assessment scenario. The question then needs to be raised as to how one could deal with the uncertainties of having more than 248 putative MC congeners (Spoof and Catherine, 2017) on one hand, yet absence of relevant toxicity data for the majority of these MC congeners, on the other hand, that would allow for appropriate hazard and risk assessment. The latter discrepancy is exacerbated by the fact that for the most of the 248 putative MC congeners there is no purified material available to actually test the MC using *in vitro* toxicokinetic and -dynamic assays and thus to provide for a minimal dataset that could be of toxicological relevance for humans.

One approach, albeit yet limited to the toxicodynamic component of the apical toxicity, is the employed ML approach for roughly categorizing MC congeners into groups of "toxic, slightly toxic, and non-toxic" MC and thus to make preliminary predictions. Although the available data for training of the ML approach was restricted to 18 MC congeners, the number of samples was artificially increased with the oversampling technique SMOTE. This approach introduced uncertainty in the data, which might have caused overlapping data points of different classes and therefore the wrong classification of compounds. Therefore, more MC should be tested, including rare MC variants (e.g. doubly-demethylated congeners) to expand the existing data set for better model performance and essentially more fine-scaled predictions (i.e. more toxicity classes). Inclusion of other PPP inhibitors, e.g. the structurally related nodularins (Rinehart et al., 1988) and structurally unrelated anabaenopeptins (Spoof et al., 2015) should ensure a better and more sensitive predictive performance. However, irrespective of the potential uncertainties experienced, both models used (Figure 5 and Table S4[1]) provided for 80-90% correct predictions of toxicity class. More importantly however, modulation of the training set allowed for improved prediction (Figure 5B), whereby most of the few wrongly predicted MC were found in a higher toxicity class, thereby overestimating the true toxicity. As overestimation of toxicity would be more precautionary with regard to potential hazard and risk, this caveat of the ML approach was considered acceptable.

Obviously categorizing MC congeners into toxicodynamic classes will not resolve the problem of having to assess the potential hazard of mixtures of different MC congeners present in a given surface water. Additionally, the use of calculated TEF, as attempted by Garibo et al. (2014) with six different MC congeners, or the TEF calculated from the PPP inhibition data in this study (Table 3), will not alleviate the problem of lacking toxicokinetic data. Although there have been past efforts to characterize the uptake into human cells via OATPs (Fischer et al., 2005, 2010; Monks et al., 2007) and current efforts are under way to characterize the efflux of MC congeners from human epithelial cells using human exporter expressing insect membrane vesicle systems (Kaur et al., 2019), we are still far from actually being able to calculate individual MC congener toxicokinetics or even considering integrating toxicokinetic- and dynamic data into toxicologically based kinetic and -dynamic modelling.

Furthermore, the dramatic differences observed for MC congener toxicity in hepatocytes derived from different human patients (Fischer et al., 2010) represent a hurdle still to overcome in the future. However, data should be revisited as i.) up to 4-fold differences are apparent just in the toxicodynamic (PPP1, 2A and 5) component of the apical toxicity (Tables 1 and 3), ii.) rodent and human OATP expression and capacity appear to differ profoundly (Fischer et al., 2005, 2010), and iii.) humans differ in OATP type and level of expression, would mandate that the original safety extrapolations (Dietrich and Hoeger, 2005; WHO, 2017; Falconer and Humpage, 2005) based on rodent. Prudence would require that the current guidance value of 1.0 $MC\text{-}LR_{equivalent}$ µg/L be lowered by at least a factor ten thus accommodating recent *in vitro* findings with human cell systems. In the future, integrated systems toxicology approaches including computational toxicology (Cronin and

Dietrich, 2017) and the ML approach presented here in combination with adverse outcome pathways[3] could provide for a much better basis for the hazard and risk assessment of MC in toxic cyanobacterial blooms and thus for better guidelines regarding the safety of surface waters used for drinking water and recreational purposes.

Tab. 3: MC congener toxicity equivalency factors (TEF)

Congener	PPP1		PPP2A		PPP5	
	IC_{50} (nM)	TEF	IC_{50} (nM)	TEF	IC_{50} (nM)	TEF
MC-RR	1.5	0.20	1.6	0.31	11.7	0.44
MC-LR	0.3	**1.00**	0.5	**1.00**	5.1	**1.00**
MC-YR	1.3	0.22	n.d.	n.d.	5.1	1.00
MC-WR	1.2	0.24	1.0	0.50	5.6	0.91
MC-LA	1.9	0.15	0.7	0.71	6.1	0.84
MC-LY	0.8	0.36	n.d.	n.d.	4.1	1.24
MC-LF	2.0	0.15	1.4	0.36	4.7	1.09
MC-LW	1.2	0.24	0.7	0.66	2.5	2.02
MC-HilR	0.6	0.49	n.d.	n.d.	4.2	1.20
MC-HtyR	0.7	0.45	n.d.	n.d.	4.7	1.09
[β-D-Asp3]-MC-RR	45.0	0.01	n.d.	n.d.	167.1	0.03
[β-D-Asp3]-MC-LR	0.9	0.34	n.d.	n.d.	10.2	0.50
[β-D-Asp3, Dhb7]-MC-RR	62.0	0.01	84.3	0.01	877.1	0.01
MC-LY(Prg)	1.7	0.17	0.4	1.76	1.7	2.95
[MSecPh7]-MC-LY(Prg)	1.9	0.15	0.9	0.54	18.2	0.28
[*enantio*-Adda5]-MC-LF	-	-	-	-	-	-
[Amba5]-MC-LY(Prg)	520817	0.000	2135	0.000	54063	0.00
[Anda5]-MC-LY(Prg)	1724	0.000	n.d.	n.d.	2420	0.00

References

Andreeva, A. V. and Kutuzov, M. A. (2001). Ppp family of protein ser/thr phosphatases: Two distinct branches? *Molecular Biology and Evolution 18*, 448-452. doi:10.1093/oxfordjournals.molbev.a003823

Asgari, E. and Mofrad, M. R. (2015). Continuous distributed representation of biological sequences for deep proteomics and genomics. *PLoS One 10*, e0141287. doi:10.1371/journal.pone.0141287

Azevedo, S. M. F. O., Carmichael, W. W., Jochimsen, E. M. et al. (2002). Human intoxication by microcystins during renal dialysis treatment in caruaru—brazil. *Toxicology 181–182*, 441-446. doi:10.1016/s0300-483x(02)00491-2

Beleites, C., Neugebauer, U., Bocklitz, T. et al. (2013). Sample size planning for classification models. *Analytica Chimica Acta 760*, 25-33. doi:10.1016/j.aca.2012.11.007

Berry, M. A., Davis, T. W., Cory, R. M. et al. (2017). Cyanobacterial harmful algal blooms are a biological disturbance to western lake erie bacterial communities. *Environ Microbiol 19*, 1149-1162. doi:10.1111/1462-2920.13640

Bowling, L. and Baker, P. D. (1996). *Major cyanobacterial bloom in the barwon-darling river, australia, in 1991, and underlying limnological conditions*. Vol. doi:10.1071/mf9960643

Brewis, N. D., Street, A. J., Prescott, A. R. et al. (1993). Ppx, a novel protein serine/threonine phosphatase localized to centrosomes. *The EMBO Journal 12*, 987-996. doi:10.1002/j.1460-2075.1993.tb05739.x

Chang, N., Vannah, B. and Yang, Y. J. (2014). Comparative sensor fusion between hyperspectral and multispectral satellite sensors for monitoring microcystin distribution in lake erie. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing 7*, 2426-2442. doi:10.1109/jstars.2014.2329913

Chawla, N. V., Bowyer, K. W., Hall, L. O. et al. (2002). Smote: Synthetic minority over-sampling technique. *Journal of artificial intelligence research 16*, 321-357. doi:10.1613/jair.953

Cohen, P. T. W. (2004). Overview of protein serine/threonine phosphatases. In J. n. Ariño and D. R. Alexander (eds.), *Protein phosphatases*. Berlin, Heidelberg: Springer Berlin Heidelberg. doi:10.1007/978-3-540-40035-6_1

Cronin, M. T. D. and Dietrich, D. (2017). Computational toxicology to support predictions of admet properties and as a promising tool for 21st century risk assessment. *Computational Toxicology 1*, 1-2. doi:10.1016/j.comtox.2017.02.001

Dellafiora, L., Dall'Asta, C. and Galaverna, G. (2018). Toxicodynamics of mycotoxins in the framework of food risk assessment-an in silico perspective. *Toxins (Basel) 10*, doi:10.3390/toxins10020052

Dietrich, D. and Hoeger, S. (2005). Guidance values for microcystins in water and cyanobacterial supplement products (blue-green algal supplements): A reasonable or misguided approach? *Toxicology and Applied Pharmacology 203*, 273-289. doi:10.1016/j.taap.2004.09.005

EFSA PPR Panel, Ockleford, C., Adriaanse, P. et al. (2018). Scientific opinion on the state of the art of toxicokinetic/toxicodynamic (tktd) effect models for regulatory risk assessment of pesticides for aquatic organisms. *EFSA Journal 16*, e05377. doi:10.2903/j.efsa.2018.5377

Falconer, I. R., Burch, M. D., Steffensen, D. A. et al. (1994). Toxicity of the blue-green alga (cyanobacterium) microcystis aeruginosa in drinking water to growing pigs, as an animal model for human injury and risk assessment. *Environmental Toxicology and Water Quality 9*, 131-139. doi:10.1002/tox.2530090209

Falconer, I. R. and Humpage, A. R. (2005). Health risk assessment of cyanobacterial (blue-green algal) toxins in drinking water. *Int J Environ Res Public Health 2*, 43-50. doi:10.3390/ijerph2005010043

[3] http://www.oecd.org/chemicalsafety/testing/projects-adverse-outcome-pathways.htm (accessed 28.06.2019)

Fawell, J. K., Mitchell, R. E., Everett, D. J. et al. (1999). The toxicity of cyanobacterial toxins in the mouse: I microcystin-lr. *Hum Exp Toxicol 18*, 162-167. doi:10.1177/096032719901800305

Feurstein, D., Holst, K., Fischer, A. et al. (2009). Oatp-associated uptake and toxicity of microcystins in primary murine whole brain cells. *Toxicology and Applied Pharmacology 234*, 247-255. doi:10.1016/j.taap.2008.10.011

Feurstein, D., Kleinteich, J., Heussner, A. H. et al. (2010). Investigation of microcystin congener–dependent uptake into primary murine neurons. *Environmental health perspectives 118*, 1370. doi:10.1289/ehp.0901289

Feurstein, D., Stemmer, K., Kleinteich, J. et al. (2011). Microcystin congener– and concentration-dependent induction of murine neuron apoptosis and neurite degeneration. *Toxicological Sciences 124*, 424-431. doi:10.1093/toxsci/kfr243

Fischer, A., Hoeger, S. J., Stemmer, K. et al. (2010). The role of organic anion transporting polypeptides (oatps/slcos) in the toxicity of different microcystin congeners in vitro: A comparison of primary human hepatocytes and oatp-transfected hek293 cells. *Toxicology and Applied Pharmacology 245*, 9-20. doi:10.1016/j.taap.2010.02.006

Fischer, W. J., Garthwaite, I., Miles, C. O. et al. (2001). Congener-independent immunoassay for microcystins and nodularins. *Environmental Science & Technology 35*, 4849-4856. doi:10.1021/es011182f

Fischer, W. J., Altheimer, S., Cattori, V. et al. (2005). Organic anion transporting polypeptides expressed in liver and brain mediate uptake of microcystin. *Toxicology and Applied Pharmacology 203*, 257-263. doi:10.1016/j.taap.2004.08.012

Fontanillo, M., Zemskov, I., Häfner, M. et al. (2016). Synthesis of highly selective submicromolar microcystin-based inhibitors of protein phosphatase (pp)2a over pp1. *Angew. Chem. 128*, 14191-14195. doi:10.1002/ange.201606449

Friedman, J. H. (2001). Greedy function approximation: A gradient boosting machine. *Ann. Statist. 29*, 1189-1232. doi:10.1214/aos/1013203451

Garibo, D., Flores, C., Ceto, X. et al. (2014). Inhibition equivalency factors for microcystin variants in recombinant and wild-type protein phosphatase 1 and 2a assays. *Environ Sci Pollut Res Int* 10652-10660. doi:10.1007/s11356-014-3065-7

Hastie, C. J., Borthwick, E. B., Morrison, L. F. et al. (2005). Inhibition of several protein phosphatases by a non-covalently interacting microcystin and a novel cyanobacterial peptide, nostocyclin. *Biochim Biophys Acta 1726*, 187-193. doi:10.1016/j.bbagen.2005.06.005

Heresztyn, T. and Nicholson, B. C. (2001). Determination of cyanobacterial hepatotoxins directly in water using a protein phosphatase inhibition assay. *Water Research 35*, 3049-3056. doi:10.1016/s0043-1354(01)00018-5

Hoeger, S. J., Schmid, D., Blom, J. F. et al. (2007). Analytical and functional characterization of microcystins [asp3]mc-rr and [asp3,dhb7]mc-rr: Consequences for risk assessment? *Environ Sci Technol 41*, 2609-2616. doi:10.1021/es062681p

Huisman, J., Codd, G. A., Paerl, H. W. et al. (2018). Cyanobacterial blooms. *Nat. Rev. Microbiol. 16*, 471-483. doi:10.1038/s41579-018-0040-1

Jaeger, S., Fulle, S. and Turk, S. (2018). Mol2vec: Unsupervised machine learning approach with chemical intuition. *J Chem Inf Model 58*, 27-35. doi:10.1021/acs.jcim.7b00616

Kaur, G., Fahrner, R., Wittmann, V. et al. (2019). Human mrp2 exports mc-lr but not the glutathione conjugate. *ChemicoBiological Interactions submitted*,

Kleinteich, J., Wood, S. A., Küpper, F. C. et al. (2012). Temperature-related changes in polar cyanobacterial mat diversity and toxin production. *Nature Climate Change 2*, 356. doi:10.1038/nclimate1418

Kleinteich, J., Puddick, J., Wood, S. A. et al. (2018). Toxic cyanobacteria in svalbard: Chemical diversity of microcystins detected using a liquid chromatography mass spectrometry precursor ion screening method. *Toxins (Basel) 10*, doi:10.3390/toxins10040147

Lemaître, G., Nogueira, F. and Aridas, C. K. (2017). Imbalanced-learn: A python toolbox to tackle the curse of imbalanced datasets in machine learning. *The Journal of Machine Learning Research 18*, 559-563.

Mackintosh, C., Beattie, K. A., Klumpp, S. et al. (1990). Cyanobacterial microcystin-lr is a potent and specific inhibitor of protein phosphatase-1 and phosphatase-2a from both mammals and higher-plants. *FEBS Lett 264*, 187-192. doi:10.1016/0014-5793(90)80245-e

Mikolov, T., Chen, K., Corrado, G. et al. (2013). Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*

Monks, N. R., Liu, S., Xu, Y. et al. (2007). Potent cytotoxicity of the phosphatase inhibitor microcystin lr and microcystin analogues in oatp1b1- and oatp1b3-expressing hela cells. *Mol Cancer Ther 6*, 587-598. doi:10.1158/1535-7163.mct-06-0500

Pedregosa, F., Varoquaux, G., Gramfort, A. et al. (2011). Scikit-learn: Machine learning in python. *J. Mach. Learn. Res. 12*, 2825-2830.

Pereira, S. R., Vasconcelos, V. M. and Antunes, A. (2011). The phosphoprotein phosphatase family of ser/thr phosphatases as principal targets of naturally occurring toxins. *Crit Rev Toxicol 41*, 83-110. doi:10.3109/10408444.2010.515564

Preece, E. P., Hardy, F. J., Moore, B. C. et al. (2017). A review of microcystin detections in estuarine and marine waters: Environmental implications and human health risk. *Harmful Algae 61*, 31-45. doi:10.1016/j.hal.2016.11.006

Puddick, J., Prinsep, M. R., Wood, S. A. et al. (2014). High levels of structural diversity observed in microcystins from microcystis cawbg11 and characterization of six new microcystin congeners. *Mar Drugs 12*, 5372-5395. doi:10.3390/md12115372

Rinehart, K. L., Harada, K., Namikoshi, M. et al. (1988). Nodularin, microcystin, and the configuration of adda. *J. Am. Chem. Soc. 110*, 8557-8558. doi:10.1021/ja00233a049

Shi, Y. (2009). Serine/threonine phosphatases: Mechanism through structure. *Cell 139*, 468-484. doi:10.1016/j.cell.2009.10.006

Spoof, L., Blaszczyk, A., Meriluoto, J. et al. (2015). Structures and activity of new anabaenopeptins produced by baltic sea cyanobacteria. *Mar Drugs 14*, 8. doi:10.3390/md14010008

Spoof, L. and Catherine, A. (2017). Appendix 3: Tables of microcystins and nodularin. In J. Meriluoto, L. Spoof and G. Codd (eds.), *Handbook of cyanobacterial monitoring and cyanotoxin analysis*. John Wiley & Sons. doi:10.1002/9781119068761.app3

Stefansson, B., Ohama, T., Daugherty, A. E. et al. (2008). Protein phosphatase 6 regulatory subunits composed of ankyrin repeat domains. *Biochemistry 47*, 1442-1451. doi:10.1021/bi7022877

Taranu, Z. E., Gregory-Eaves, I., Steele, R. J. et al. (2017). Predicting microcystin concentrations in lakes and reservoirs at a continental scale: A new framework for modelling an important health risk factor. *Global Ecology and Biogeography 26*, 625-637. doi:10.1111/geb.12569

WHO - World Health Organization (1999). *Toxic cyanobacteria in water: A guide to their public health consequences monitoring and management*. Vol. E & FN Spon. doi:10.1201/9781482295061

WHO (2017). *Guidelines for drinking-water quality*. Vol. 4. Geneva: World Health Organization. doi:10.1007/springerreference_30502

Wood, S. A., Rueckert, A., Hamilton, D. P. et al. (2011). Switching toxin production on and off: Intermittent microcystin synthesis in a microcystis bloom. *Environ Microbiol Rep 3*, 118-124. doi:10.1111/j.1758-2229.2010.00196.x

Wu, Z., Ramsundar, B., Feinberg, E. N. et al. (2018). Moleculenet: A benchmark for molecular machine learning. *Chem Sci 9*, 513-530. doi:10.1039/c7sc02664a

Yang, J., Roe, S. M., Cliff, M. J. et al. (2005). Molecular basis for tpr domain-mediated regulation of protein phosphatase 5. *The EMBO Journal 24*, 1-10. doi:10.1038/sj.emboj.7600496

Zemskov, I., Altaner, S., Dietrich, D. R. et al. (2017). Total synthesis of microcystin-lf and derivatives thereof. *J Org Chem 82*, 3680-3691. doi:10.1021/acs.joc.7b00175

Conflict of interest

The authors declare that they have no conflicts of interest.

Acknowledgments

We gratefully acknowledge the Arthur-und-Aenne-Feindt foundation (Hamburg, Germany), the Konstanz Research School Chemical Biology (KoRS-CB) and CHARM (BadenWürttemberg Wassernetzwerk) for financial support. The UPLC-MS/MS was funded through a large investment grant of the DFG INST 38-537-1.

Author contributions

Study was designed by S. Altaner, D.R. Dietrich, S. Jaeger and F. Schreiber. Experiments for PPP and PPP2a were performed by S. Altaner, while PPP5 experiments were performed by R. Fotler under the supervision of S. Altaner (Bachelor-thesis R. Fotler). Synthetic microcystin variants were produced by I. Zemskov under the supervision of V. Wittmann (PhD Thesis I. Zemskov). Machine learning and prediction models were performed by S. Jaeger. Manuscript was written by S. Altaner, S. Jaeger (machine learning part) and D.R. Dietrich. All authors corrected, amended and complemented the manuscript.