Bioinformatics, 32, 2016, i28–i36
doi: 10.1093/bioinformatics/btw246
ISMB 2016

OXFORD

Fast metabolite identification with Input Output Kernel Regression

Céline Brouard,[1,2,*] Huibin Shen,[1,2] Kai Dührkop,[3] Florence d'Alché-Buc,[4] Sebastian Böcker[3] and Juho Rousu[1,2]

[1]Department of Computer Science, Aalto University, Espoo, Finland, [2]Helsinki Institute for Information Technology, Espoo, Finland, [3]Chair for Bioinformatics, Friedrich-Schiller University, Jena, Germany and [4]LTCI, CNRS, Télécom ParisTech, Université Paris-Saclay, Paris, France

*To whom correspondence should be addressed.

Abstract

Motivation: An important problematic of metabolomics is to identify metabolites using tandem mass spectrometry data. Machine learning methods have been proposed recently to solve this problem by predicting molecular fingerprint vectors and matching these fingerprints against existing molecular structure databases. In this work we propose to address the metabolite identification problem using a structured output prediction approach. This type of approach is not limited to vector output space and can handle structured output space such as the molecule space.

Results: We use the Input Output Kernel Regression method to learn the mapping between tandem mass spectra and molecular structures. The principle of this method is to encode the similarities in the input (spectra) space and the similarities in the output (molecule) space using two kernel functions. This method approximates the spectra-molecule mapping in two phases. The first phase corresponds to a regression problem from the input space to the feature space associated to the output kernel. The second phase is a preimage problem, consisting in mapping back the predicted output feature vectors to the molecule space. We show that our approach achieves state-of-the-art accuracy in metabolite identification. Moreover, our method has the advantage of decreasing the running times for the training step and the test step by several orders of magnitude over the preceding methods.

Availability and implementation:

Contact: celine.brouard@aalto.fi

Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 Introduction

Metabolomics is a science which concerns the study of small molecules, called metabolites, and their interactions in the cell. An important problem of metabolomics is the identification of the metabolites present in a sample. Information on metabolites can be obtained using tandem mass spectrometry. This technology allows to obtain a tandem mass spectrum, also called MS/MS spectrum, by fragmenting a compound. A MS/MS spectrum is a plot containing a set of peaks, where each peak corresponds to a fragment. These peaks represent the relative abundance of the different fragments, also called intensity, in function of their mass-to-charge ratio. The identification of the metabolite from its mass spectrum is then needed for a more detailed biological interpretation. In general this step consists in a research of the obtained spectrum in databases of reference spectra, followed by an analysis by experts of the domain.

Computational approaches for interpreting and predicting MS/MS data of small molecules date back to the 1960s (Lindsay *et al.*, 1980). However, the early approaches were hampered by the unavailability of large scale data on molecular structures as well as reference spectra. The introduction of molecular structure databases such as PubChem (Bolton *et al.*, 2008) as well as open mass spectral reference databases (da Silva *et al.*, 2015; Horai *et al.*, 2010) has in recent years fuelled the development of novel methods. Several novel strategies have been proposed, including simulation of mass spectra from molecular structure (Allen *et al.*, 2014, 2015), combinatorial fragmentation (Heinonen *et al.*, 2008; Hill and Mortishire-Smith, 2005; Ridder *et al.*, 2013; Wang *et al.*, 2014; Wolf *et al.*, 2010) and prediction of molecular fingerprints (Heinonen *et al.*, 2012; Shen *et al.*, 2014).

Methods based on machine learning (Allen *et al.*, 2014, 2015; Dührkop *et al.*, 2015; Heinonen *et al.*, 2012; Shen *et al.*, 2013,

2014) have been proposed very recently for learning a mapping between tandem mass spectra and metabolites. These methods fall into two general approaches. The first group of methods (Dührkop *et al.*, 2015; Heinonen *et al.*, 2012; Shen *et al.*, 2013, 2014) introduces an intermediary step consisting in predicting molecular fingerprints for the metabolites from their mass spectra using Support Vector Machines (SVMs).

Molecular fingerprints are a standard representation for molecules, used in cheminformatics and drug discovery. They are typically represented as binary vectors, whose values indicate the presence or absence of some molecular properties, e.g. the existence of particular substructures in the metabolite or some physiochemical properties. If two molecules share a large number of molecular properties they are likely to be similar in structure, which is the rationale in using them for metabolite identification. To identify a metabolite, the fingerprint predicted from its tandem mass spectrum is matched against a large molecular database such as PubChem. In Shen *et al.* (2014) and Dührkop *et al.* (2015) fragmentation trees are computed to model the fragmentation process of the molecules and then used for predicting the molecular fingerprints. The other machine learning approach for metabolite identification, used by CFM-ID (Allen *et al.*, 2014, 2015), also relies on a two-step scheme, where the first step consists in predicting the mass spectra of the candidate molecules by modeling their fragmentation processes. In the second step, the simulated spectra of the candidate molecules are compared with the spectrum of the test metabolite.

The goal of this work is to solve the metabolite identification problem in a single step, using a structured prediction method. These methods make use of structural dependencies existing among complex outputs (e.g. the fingerprints of a molecule) to improve the accuracy and make prediction efficiently. These methods have achieved an improved prediction performance over methods that predict parts of a structure independently in numerous applications. In the literature, two main structured prediction approaches can be distinguished. The first one models the dependencies between structured inputs and outputs using a joint feature map $\phi(x, y)$ (Marchand *et al.*, 2014; Rousu *et al.*, 2007; Su and Rousu, 2015; Taskar *et al.*, 2004; Tsochantaridis *et al.*, 2004), and learns to discriminate the correct structure y for an input x from all incorrect output structures. The second one, called Output Kernel Regression, consists in learning a mapping between the input set and the feature space associated to some output kernel. A preimage problem, which consists in mapping back the predicted output feature vectors to the output space, is then solved. Existing Output Kernel Regression methods are Kernel Dependency Estimation (Cortes *et al.*, 2005; Kadri *et al.*, 2013; Weston *et al.*, 2003), Output Kernel Trees (Geurts *et al.*, 2006) and Input Output Kernel Regression (IOKR) (Brouard *et al.*, 2011, 2015).

In this work, we show how to apply the IOKR framework for solving the metabolite identification problem. Our method reaches improved identification rates compared with the previous state-of-the-art of Dührkop *et al.* (2015). More importantly, though, the IOKR framework results in vast improvements in running times: the method is one to two orders of magnitude faster in the prediction phase, and four orders of magnitude faster during training.

2 Methods

The main notations used in this article are summarized in Table 1. In the following, we note $\mathcal{X}$ the set of input tandem mass spectra, also known as MS/MS spectra, and $\mathcal{Y}$ the set containing the 2D molecular structures corresponding to the spectra. We want to learn a function f that maps a MS/MS spectrum $x \in \mathcal{X}$ to its corresponding molecular structure $y \in \mathcal{Y}$. In this problem both input and output data are structured. Structured data refer to data having an internal structure, for example a graph or a tree, or to data being interdependent to each other. To solve this problem we use the IOKR framework that can learn a mapping between structured inputs and structured outputs. This framework has been introduced by Brouard *et al.* (2011) to solve link prediction in the semi-supervised setting. In Brouard *et al.* (2015), this approach has been extended to address general structured output prediction problems. In this section we describe this method and explain how it can be applied to solve metabolite identification.

Table 1. Notations used in the article

Symbol	Explanation
$\mathcal{X}$, $\mathcal{Y}$	input, output sets
x, y	elements of $\mathcal{X}$, $\mathcal{Y}$
$\kappa_y : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}$	output scalar kernel
$\mathcal{F}_y$	output feature space
$\phi_y : \mathcal{Y} \to \mathcal{F}_y$	output feature map
$\mathcal{K}_x : \mathcal{X} \times \mathcal{X} \to \mathcal{L}(\mathcal{F}_y, \mathcal{F}_y)$	input operator-valued kernel
$\mathcal{H}$	reproducing kernel Hilbert space of $\mathcal{K}_x$
$\mathbf{K}_{X_\ell}$	Gram matrix on training set
$\kappa_x : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$	input scalar kernel
$\mathcal{F}_x$	input feature space
$\phi_x : \mathcal{X} \to \mathcal{F}_x$	input feature map
K_{X_ℓ}	Gram matrix on training set

In the IOKR approach the internal structure of the output data is encoded using a kernel function $\kappa_y : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}$. A kernel function is a positive semi-definite function that measures the similarity between two elements. Its values can be evaluated by computing scalar products in a high-dimensional space, called the feature space. In the case of the output kernel κ_y, this writes as follows:

$$\forall (y, y') \in \mathcal{Y} \times \mathcal{Y},\ \kappa_y(y, y') = \langle \phi_y(y), \phi_y(y') \rangle_{\mathcal{F}_y},$$

where the Hilbert space $\mathcal{F}_y$ is the feature space associated to κ_y and $\phi_y : \mathcal{Y} \to \mathcal{F}_y$ is a feature map that maps the outputs to the output feature space. Depending of the kernel used, for example when using a Gaussian kernel, the feature map ϕ_y might not be explicitly known. We will see later that we only need to evaluate inner products between feature vectors for computing the solution, which is possible using the kernel trick in the output space. This means that the scalar products in the feature space are replaced by the kernel values.

The spectra-metabolite mapping problem can then be decomposed in two tasks (see Figure 1). The first task consists in learning a function h between the input set $\mathcal{X}$ and the Hilbert space $\mathcal{F}_y$ that approximates the feature map ϕ_y. This task is called *Output Kernel Regression*. The second task is a pre-image problem that requires to learn or define a function g from $\mathcal{F}_y$ to the output set $\mathcal{Y}$. We detail these two steps in the following subsections.

2.1 Output Kernel Regression

The values of the function h that we want to learn in the Output Kernel Regression step are vectors belonging to the Hilbert space $\mathcal{F}_y$ and not scalars. IOKR uses the Reproducing Kernel Hilbert Space (RKHS) theory devoted to vector-valued functions (Micchelli and

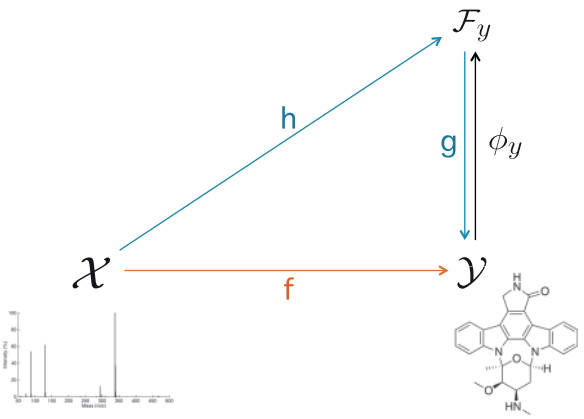

Fig. 1. Overview of the IOKR framework for solving the metabolite identification problem. The mapping f between MS/MS spectra and 2D molecular structures is learnt by approximating the output feature map ϕ_y with a function h and solving a preimage problem

Pontil, 2005; Senkene and Tempel'man, 1973) in order to find an appropriate functional space $\mathcal{H}$ for searching the function h. This theory extends nicely the kernel methods to the problem of learning vector-valued functions. It has been used in the literature to solve different learning problems such as multi-task learning (Evgeniou *et al.*, 2005), functional regression (Kadri *et al.*, 2010), link prediction (Brouard *et al.*, 2011) and vector autoregression (Lim *et al.*, 2014).

In this theory, a kernel $\mathcal{K}_x : \mathcal{X} \times \mathcal{X} \to \mathcal{L}(\mathcal{F}_y, \mathcal{F}_y)$ is a function whose values are linear operators from $\mathcal{F}_y$ to $\mathcal{F}_y$, where $\mathcal{F}_y$ is a general Hilbert space. This theory does not require any assumption on the existence of an output kernel κ_y. $\mathcal{K}_x$ is called an operator-valued kernel if it verifies the two following properties:

1. $\forall x, x' \in \mathcal{X},\ \mathcal{K}_x(x, x') = \mathcal{K}_x(x', x)^*$, where $*$ denotes the adjoint. $\mathcal{K}_x(x', x)^*$ is defined as the linear operator satisfying $\langle \mathcal{K}_x(x', x)\tilde{\mathbf{y}}_i, \tilde{\mathbf{y}}_j\rangle_{\mathcal{F}_y} = \langle \tilde{\mathbf{y}}_i, \mathcal{K}_x(x', x)^*\tilde{\mathbf{y}}_j\rangle_{\mathcal{F}_y},\ \forall \tilde{\mathbf{y}}_i, \tilde{\mathbf{y}}_j \in \mathcal{F}_y$
2. $\forall m \in \mathbb{N},\ \forall \{(x_i, \tilde{\mathbf{y}}_i)\}_{i=1}^m \subseteq \mathcal{X} \times \mathcal{F}_y, \sum_{i,j=1}^m \langle \tilde{\mathbf{y}}_i, \mathcal{K}_x(x_i, x_j)\tilde{\mathbf{y}}_j\rangle_{\mathcal{F}_y} \geq 0$

In the case where the dimension d of $\mathcal{F}_y$ is finite, the kernel $\mathcal{K}_x$ is a function whose values are matrices of size $d \times d$ and the kernel matrix is a block matrix.

In the IOKR approach, the function $h : \mathcal{X} \to \mathcal{F}_y$ is searched in the RKHS with reproducing kernel $\mathcal{K}_x$. We denote this space $\mathcal{H}$. This means that we are searching models of the following form:

$$\forall x \in \mathcal{X},\ h(x) = \sum\nolimits_i \mathcal{K}_x(x, x_i)\mathbf{c}_i,\ \ \mathbf{c}_i \in \mathcal{F}_y.$$

Let $\{(x_i, \phi_y(y_i))\}_{i=1}^\ell \subseteq \mathcal{X} \times \mathcal{F}_y$ be the set of training examples. The function h is searched by minimizing a regularized optimization problem. In this article, we chose to use the regularized least-squares loss function in the supervised setting:

$$\underset{h \in \mathcal{H}}{\operatorname{argmin}} \sum\nolimits_{i=1}^\ell ||h(x_i) - \phi_y(y_i)||^2_{\mathcal{F}_y} + \lambda \|h\|^2_{\mathcal{H}}, \tag{1}$$

where $\lambda > 0$ is a regularization parameter. A sufficiently high enough value of λ prevents overfitting. According to the Representer Theorem (Micchelli and Pontil, 2005), the solution of this optimization problem can be written as a linear combination of the operator-valued kernel evaluated on the training examples:

$$\widehat{h}(x_i) = \sum\nolimits_{j=1}^\ell \mathcal{K}_x(x_i, x_j)\mathbf{c}_j,$$

where $\mathbf{c}_j$, $j = 1, \ldots, \ell$, are vectors in $\mathcal{F}_y$. By replacing this expression in the optimization problem (1) and computing the derivative of the optimization problem, it has been shown by Micchelli and Pontil (2005) that the vectors $\mathbf{c}_j$, $j = 1, \ldots, \ell$ verify the following equation:

$$\sum\nolimits_{i=1}^\ell (\mathcal{K}_x(x_j, x_i) + \lambda\delta_{ij})\mathbf{c}_i = \phi_y(y_j),$$

where $\delta_{ii} = 1$ and $\delta_{ij} = 0$ for all $j \neq i$.

If the dimension d of the output feature space $\mathcal{F}_y$ is finite, this solution can be rewritten in closed form as follows:

$$\mathrm{vec}(C_\ell) = (\lambda I_{\ell d} + \mathbf{K}_{X_\ell})^{-1}\mathrm{vec}(\Phi_{Y_\ell}), \tag{2}$$

where $C_\ell = (\mathbf{c}_1, \ldots, \mathbf{c}_\ell)$ and $\Phi_{Y_\ell} = (\phi_y(y_1), \ldots, \phi_y(y_\ell))$ are two matrices of size $d \times \ell$; $I_{\ell d}$ denotes the identity matrix of size $\ell d \times \ell d$; and $\mathbf{K}_{X_\ell}$ is the Gram matrix of the operator-valued kernel $\mathcal{K}_x$ on the training set. This is a $\ell \times \ell$ block matrix, each block being of size $d \times d$. $\mathrm{vec}(C_\ell)$ is a column vector of length ℓd obtained by stacking the columns of the matrix C_ℓ on top of each other. Equation (2) generalizes the solution obtained with kernel ridge regression to the case of vector-valued functions.

2.2 Preimage step

To predict the output metabolite $f(x)$ associated to the spectra $x \in \mathcal{X}$, we must determine the pre-image of $h(x)$ by ϕ_y. For this, we search the metabolite y in a set of candidates $\mathcal{Y}^*$ that minimizes the following criteria:

$$\widehat{f}(x) = \underset{y \in \mathcal{Y}^*}{\operatorname{argmin}} ||\widehat{h}(x) - \phi_y(y)||^2_{\mathcal{F}_y}. \tag{3}$$

As we consider that the output kernel is normalized, Equation (3) becomes:

$$\widehat{f}(x) = \underset{y \in \mathcal{Y}^*}{\operatorname{argmax}} \langle \widehat{h}(x), \phi_y(y)\rangle_{\mathcal{F}_y}.$$

In this work, we consider operator-valued kernels of the following form:

$$\forall (x, x') \in \mathcal{X} \times \mathcal{X},\ \mathcal{K}_x(x, x') = \kappa_x(x, x') * I_d, \tag{4}$$

where $\kappa_x : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ is a scalar input kernel. We note $\mathcal{F}_x$ the Hilbert space associated to this kernel and $\phi_x : \mathcal{X} \to \mathcal{F}_x$ a feature map of κ_x. By using this operator-valued kernel and replacing $\widehat{h}$ by the solution given in the previous subsection, we obtain the following solution for metabolite identification with IOKR:

$$\widehat{f}(x) = \underset{y \in \mathcal{Y}^*}{\operatorname{argmax}}\ \phi_y(y)^{\mathrm{T}}\Phi_{Y_\ell}(\lambda I_\ell + K_{X_\ell})^{-1}\Phi_{X_\ell}^{\mathrm{T}}\phi_x(x),$$

where $\Phi_{X_\ell} = (\phi_x(x_1), \ldots, \phi_x(x_\ell))$ and K_{X_ℓ} is the Gram matrix of the scalar kernel κ_x on the training set. Using the kernel trick in the output space allows us to evaluate $\widehat{f}(x)$ even in the case where the output feature map ϕ_y is not known explicitly. The solution can be rewritten as follows:

$$\widehat{f}(x) = \underset{y \in \mathcal{Y}^*}{\operatorname{argmax}}\ (\mathbf{k}^y_{Y_\ell})^{\mathrm{T}}(\lambda I_\ell + K_{X_\ell})^{-1}\mathbf{k}^x_{X_\ell},$$

where $\mathbf{k}^y_{Y_\ell} = \begin{bmatrix} \kappa_y(y_1, y) \\ \ldots \\ \kappa_y(y_\ell, y) \end{bmatrix}$ and $\mathbf{k}^x_{X_\ell} = \begin{bmatrix} \kappa_x(x_1, x) \\ \ldots \\ \kappa_x(x_\ell, x) \end{bmatrix}$ are two column vectors.

2.3 Kernels

In the following, we describe the pairs of kernels (κ_y, κ_x) that we used for solving the metabolite identification problem with IOKR.

2.3.1 Input kernels

We considered several existing mass spectral kernels for the scalar input kernel κ_x (Dührkop *et al.*, 2015; Heinonen *et al.*, 2012; Shen *et al.*, 2014). The kernels we used in this article are listed in Table 2. Most of them are defined based on *fragmentation trees* (Dührkop *et al.*, 2015; Shen *et al.*, 2014). Introduced by Böcker and Rasche (2008), fragmentation trees model the fragmentation process of a molecule in a tree shape: nodes of this tree are molecular formulas that correspond to the unfragmented molecule and its fragments. An edge between two nodes indicates the existence of a fragmentation reaction between two fragments or between the unfragmented molecule and one of its fragments. These edges are directed and correspond to losses. An example of fragmentation tree is given in Figure 2. Based on fragmentation trees, different categories of kernels have been proposed, such as: loss-based kernels, node-based kernels, path-based kernels or fragmentation tree alignment kernels.

We also used the recalibrated probability product kernel (PPKr), which is computed on preprocessed spectra. The PPK kernel, introduced by Heinonen *et al.* (2012), is computed from MS/MS spectra by modeling each peak in a spectrum by a normal distribution with two dimensions: the mass-to-charge ratio and the intensity. A spectrum is then modeled as a mixture of normal distributions. The PPK kernel between two spectra is evaluated by integrating the product between the two corresponding mixture distributions.

We learned a linear combination of these 24 input kernels using multiple kernel learning (MKL). We used uniform MKL

Table 2. Description of the input kernels used in this article

Category	Name	Description	Reference
Loss-based kernels	Loss binary (LB)	counts the number of common losses	Shen et al. (2014)
	Loss intensity (LI)	weighted variant of LB that uses the intensity of terminal nodes	Shen et al. (2014)
	Loss count (LC)	counts the number of occurrences of the losses	Shen et al. (2014)
	Weighted loss count (LW)	weighted variant of LC using the inverse frequency of training losses	
	Root loss binary (RLB)	counts the number of common losses from the root to some node	Shen et al. (2014)
	Root loss intensity (RLI)	weighted variant of RLB that uses the intensity of terminal nodes	Shen et al. (2014)
	Loss intensity PP (LIPP)	probability product (PP) of shared losses	Dührkop et al. (2015)
Node-based kernels	Node binary (NB)	counts the number of nodes with the same molecular formula	Shen et al. (2014)
	Node intensity (NI)	weighted variant of NB that uses the intensity of nodes	Shen et al. (2014)
	Node subformula (NSF)	counts the number of common substructures	Dührkop et al. (2015)
	Fragment intensity PP (FIPP)	PP of shared fragments (nodes)	Dührkop et al. (2015)
Path-based kernels	Common paths counting (CPC)	counts the number of common paths (identical sequences of losses)	Shen et al. (2014)
	Common paths of length 2 (CP2)	counts the number of common paths of length 2	Shen et al. (2014)
	Common paths of length at least 2 (CP2+)	counts the number of common paths of length at least 2	Dührkop et al. (2015)
	Common paths with K_{peaks} (CPK1)	the PPK K_{peaks} are used to score the terminal peaks	Shen et al. (2014)
	Common paths with K_{peaks} (CPK2)	same as CPK1 with a different parameter	Shen et al. (2014)
	Common path joined binary (CPJB)	counts the number of paths for which the union of losses is equal	Dührkop et al. (2015)
	Common path joined (CPJ)	counts paths of length 2 that have the same loss	
	Weighted paths counting (WPC)	weighted variant of CPC that uses the inverse frequency of the losses	
Subtree kernel	Common subtree counting (CSC)	counts the number of subtrees with common structures and losses	Shen et al. (2014)
Fragmentation tree alignment kernels	TALIGN	Pearson correlation of alignment scores between fragmentation trees	Dührkop et al. (2015)
	TALIGND	variant of TALIGN that modifies the scoring function	Dührkop et al. (2015)
Probability product kernel	Recalibrated PPK (PPKr)	PPK computed on preprocessed spectra	Dührkop et al. (2015) Heinonen et al. (2012)
other	Chemical element counting (CEC)	weighted counts of chemical elements	Dührkop et al. (2015)

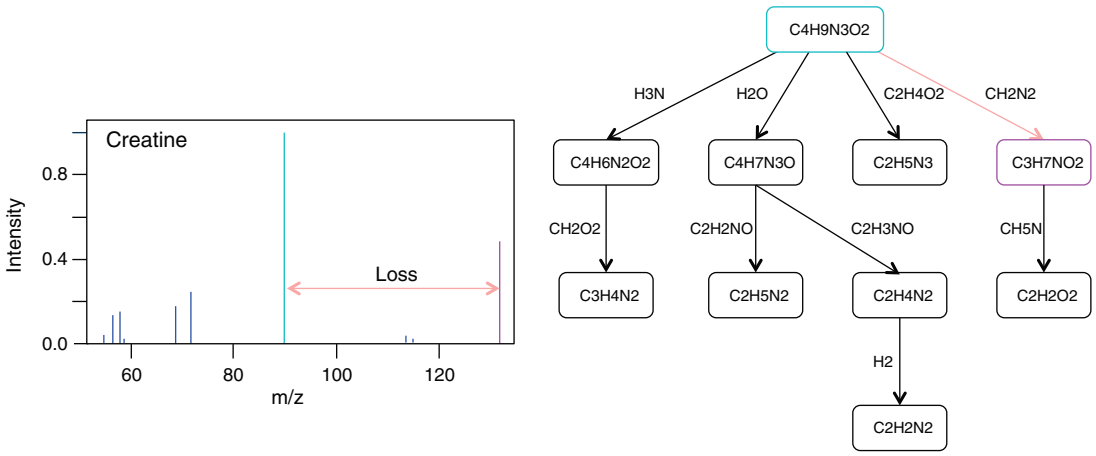

Fig. 2. An example of MS/MS spectrum and its fragmentation tree. Each node of the fragmentation tree corresponds to a peak and is labeled by the molecular formula of the corresponding fragment. The root of the tree is labeled with the molecular formula of the unfragmented molecule. Edges represent the losses. Two nodes and one edge are colored to show the correspondence between the MS/MS spectrum and the fragmentation tree

(UNIMKL), which associates the same weight to each kernel. We also applied the ALIGNF approach (Cortes *et al.*, 2012) which obtained the best performance for metabolite identification in the comparison performed by Shen *et al.* (2014). ALIGNF searches to maximize the centered kernel alignment between the combined kernel matrix and an ideal target kernel matrix K_y:

$$\max_{\mu \geq 0, ||\mu||_2 = 1} \frac{\langle \sum_{k=1}^{m} \mu_k K_k^c, K_y \rangle_{\mathrm{F}}}{||\sum_{k=1}^{m} \mu_k K_k^c||_{\mathrm{F}}}.$$

K_k^c denotes the centered Gram matrices of the input kernels:

$$K_k^c = \left[I_\ell - \frac{\mathbf{1}\mathbf{1}^{\mathrm{T}}}{\ell} \right] K_k \left[I_\ell - \frac{\mathbf{1}\mathbf{1}^{\mathrm{T}}}{\ell} \right],$$

where $\mathbf{1}$ is a column vector of ones of length ℓ. In Cortes *et al.* (2012), the target kernel was defined as $K_y = \mathbf{y}^{\mathrm{T}}\mathbf{y}$ in the case of single label classification. Here we used the Gram matrix of the output kernel κ_y on the training set. The combination of kernels learned with ALIGNF was then used for the scalar input kernel κ_x in IOKR.

2.3.2 Output kernels

For the output kernel, we have to define a similarity that takes into account the inherent structure of the metabolites. We compared the results obtained using different graph kernels (path, shortest-path and graphlet kernels) as well as kernels defined on molecular fingerprints. A molecular fingerprint is a vector encoding the structure of a molecule. Generally the values of this vector are binary values that indicate the presence or absence of certain molecular properties. A bit can indicate for example the presence of a chemical atom, a type of ring, an atom pair or a common functional group in the structure of the molecule.

We consider here the kernels that obtained the best performances, which are the kernels based on fingerprints. We used the set of 2,765 binary molecular properties described in Dührkop *et al.* (2015). More details about these molecular properties are given in the Supplementary Materials. In the experiments, we considered different type of output kernels:

- linear kernel: $\kappa_y(y, y') = c(y)^{\mathrm{T}} c(y')$,
- polynomial kernel: $\kappa_y(y, y') = \left(c(y)^{\mathrm{T}} c(y') + a \right)^b$,
- Gaussian kernel: $\kappa_y(y, y') = \exp(-\gamma ||c(y) - c(y')||^2)$,

where $c(y)$ and $c(y')$ denote the molecular fingerprints of y and y'.

3 Results

We evaluated and compared our approach on a subset of 4138 MS/MS spectra extracted from the GNPS (Global Natural Products Social) public spectral library (https://gnps.ucsd.edu/ProteoSAFe/libraries.jsp) in Dührkop *et al.* (2015).

3.1 Protocol

The evaluation was performed using a 10-fold cross-validation (10-CV) procedure such that all compounds having the same structure are contained in the same fold. The input and output kernels were centered and normalized. The regularization parameter λ and the parameter(s) of the output kernel were selected using leave-one-out CV on each training fold. We used the averaged mean squared error (MSE) as error measure for tuning these parameters. The leave-one-out estimate of the averaged MSE was computed using the closed-form solution proved in Brouard *et al.* (2015).

In the prediction step, the method was evaluated on 3,868 compounds. For solving the pre-image step, following Dührkop *et al.* (2015) we assumed that all spectra have already their molecular formula predicted as a preprocessing step, and we searched among the PubChem (Bolton *et al.*, 2008) structures having the same molecular formula as the current target. We computed the distance between the predicted output feature vector $\widehat{h}(x)$ (see Equation 3) and the output feature vectors of all the candidates. After the pre-image step, we ranked the candidates according to their distances to $\widehat{h}(x)$ (from the smallest distance value to the highest one). For the evaluation, we evaluated the rank obtained by the true molecular structure among the candidate set for each test example and then we computed the percentage of structures that have been ranked lower than k, and this for varying k values. A test compound is said to be correctly identified if its correct structure is ranked first in the list.

3.2 Comparison with competing methods

We compared the performances of our method with two competing methods: FingerID (Heinonen *et al.*, 2012) and CSI:FingerID. Dührkop *et al.* (2015) showed that CSI:FingerID improved significantly the metabolite identification rate compared with competing methods including CFM-ID (Allen *et al.*, 2015), MetFrag (Wolf *et al.*, 2010), MAGMa (Ridder *et al.*, 2013), MIDAS (Wang *et al.*, 2014) as well as FingerID—the second most accurate method in their comparison. Both FingerID and CSI:FingerID train a SVM classifier for each molecular property. A scoring function is then used to compare the predicted fingerprint with the candidate fingerprints and the candidate fingerprints are sorted correspondingly. FingerID uses as input the PPK kernel, whereas CSI:FingerID learns a combination of this kernel with different kernels defined on fragmentation trees using ALIGNF. In our experiment, we evaluated the performances of CSI:FingerID with unit scoring and with the modified Platt score, which was shown to perform the best among the different scores compared by Dührkop *et al.* (2015).

3.2.1 Identification performance

CSI:FingerID and FingerID were retrained on the 4138 GNPS spectra. For all methods, the parameter(s) were tuned on the training set using an internal 10-CV procedure. For the SVM-based approaches, the soft margin parameter C was tuned independently for each SVM. Table 3 shows the results obtained with IOKR, FingerID and CSI:FingerID and the differences with the identification percentage of CSI:FingerID modified Platt are visualized in Figure 3. We observe that IOKR with UNIMKL combined kernel and Gaussian output kernel reaches the first position with 30.66% of correct identifications that are ranked first. It is followed by IOKR linear UNIMKL, IOKR Gaussian ALIGNF and then by CSI:FingerID modified Platt with 28.84% of correctly identified metabolites. When considering the identification percentage between top 1 and

Table 3. Comparison of the percentage of correctly identified structures for top 1, 10 and 20 using FingerID, CSI:FingerId and IOKR

Method	MKL	Top 1	Top 10	Top 20
FingerID	none	17.74	49.59	58.17
CSI:FingerID unit	ALIGNF	24.82	60.47	68.2
CSI:FingerID mod Platt	ALIGNF	28.84	66.07	73.07
IOKR linear	ALIGNF	28.54	65.77	73.19
	UNIMKL	30.02	66.05	73.66
IOKR Gaussian	ALIGNF	29.78	67.84	74.79
	UNIMKL	**30.66**	**67.94**	**75.00**

The highest values are shown in boldface.

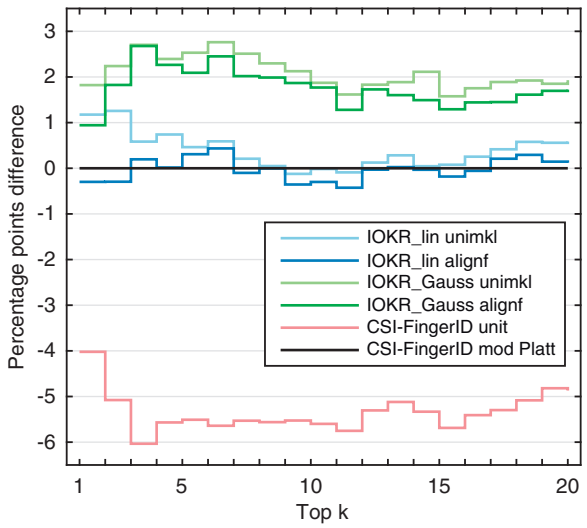

Fig. 3. Difference in percentage points to the percentage of metabolites ranked lower than k with CSI:FingerID using the modified Platt scoring function

top 20, we observe that IOKR outperforms CSI:FingerID unit in all the cases. When using a Gaussian kernel in output, IOKR improves upon CSI:FingerID modified Platt by around 2 percentage units. We performed statistical significance tests of the identification performance for the different approaches. These tests show that the difference between CSI:FingerID modified Platt and IOKR using a Gaussian output kernel is very significant. The corresponding P-values are 1.8e-16 with UNIMKL combination and 8e-14 with ALIGNF combination. A table containing all the P-values is given in the Supplementary Materials.

3.2.2 Running times

We computed the running times of CSI:FingerID and IOKR using the 4138 spectra from GNPS as training set and 625 spectra from the Massbank dataset (Horai *et al.*, 2010) as test set (see Table 4). The running times correspond to the times that would have been obtained if we were using a single core. The training times were computed using fixed values for the parameters (regularization and kernel parameters). The computation of the fragmentation trees, input kernels and fingerprints was not taken into account here. The running times for the training and the test steps are shown in Table 4. In this table, we observe a substantial difference between the training times obtained with these two approaches: the IOKR method is approximately 7000 times faster to train. This can be explained by the fact that CSI:FingerID needs to train a SVM classifier for each molecular property, this means 2765 SVMs to train in this experiment. For the same reason, IOKR also presents smaller test time compared with CSI:FingerID. In the case of the linear kernel, the test running time of IOKR is smaller than when using a Gaussian or polynomial kernel. This comes from the fact that we can avoid kernel computations in the pre-image step for the linear kernel by computing explicitly the output feature vectors.

3.3 Detailed evaluation of identification with IOKR

We will now analyze more in details the results obtained with our method on the GNPS dataset.

We begin by presenting the results obtained for the different input and output kernels introduced in Section 2. Figure 4 contains

Table 4. Running time evaluation

	Training time	Test time
CSI:FingerID	82 h 28 min 23 s	1 h 11 min 31 s
IOKR linear	**42 s**	**1 min 15 s**
IOKR polynomial	**38 s**	21 min 58 s
IOKR Gaussian	**41 s**	33 min 15 s

These running times were obtained by training the methods on the 4138 GNPS spectra and using 625 spectra from Massbank as test set.

the percentage of correctly identified structures (i.e. correct structures ranked top over all candidates) obtained with IOKR for the different pairs of input and output kernels. The two last columns correspond to the linear kernel combinations with UNIMKL and ALIGNF. We observe that the two MKL approaches clearly improve the results compared with the single kernels. The best performance is obtained with the UNIMKL approach, which is performing slightly better than ALIGNF. 30.74% of the metabolites are correctly identified with UNIMKL combined kernel. Among the individual input kernels, tree alignment-based kernels, node-based kernels [except Node subformula (NSF)] and the PPKr kernel obtain the best results. At the opposite end, the loss-based kernels and chemical element counting (CEC) are associated with low percentage of correct identified metabolites. Regarding output kernels, we notice that the performance obtained with linear and polynomial kernels are the same. This is because the optimal parameters selected for the polynomial kernel are 0 for the offset parameter and 1 for the degree, thus equalling linear kernel. Using Gaussian kernel seems to slightly improve the percentage of correctly identified structures for some input kernels, except for the root loss binary (RLB) kernel.

The averaged kernel weights learned with the ALIGNF algorithm on the training folds are visualized in Figure 5 for the three output kernels. The PPKr kernel is selected with the highest weight by ALIGNF for the three output kernels. Consistently with Figure 4, linear and the polynomial kernels are effectively the same. We observe that the weights are quite sparse: 14 kernels on a total of 24 are associated to a weight that is lower than 10^{-6}. In order to analyze why these 10 particular kernels are selected by ALIGNF, we plotted the pairwise kernel alignment scores between the input kernels, as well as the alignment scores between the input and output kernels (see in the Supplementary Materials). The first plot shows which input kernels are similar to each other. Nine groups of kernels can be distinguished and we notice that at least one kernel in each group is selected by ALIGNF. The only exception is the group containing the subtree kernel CSC but this might be because this input kernel is the one having the lowest alignment score with the output kernel. The sparsity of the kernel weights can therefore be explained by the fact that some kernels are very similar to each other and thus contain redundant information.

3.4 Prediction analysis

In the following, we detail the performance of the testing metabolites with IOKR in function of the size of their candidate sets. For this, we consider the best pair of kernels: UNIMKL combined kernel in input and Gaussian kernel in output. Figure 6a shows the distribution of the sizes of candidate sets, and the figure 6b represents the percentage of correctly identified metabolites in top 1, top 10 and above. We observe that the majority of the candidate sets contain $<$1000 candidates in our dataset. For these candidate sets, 32.8% of metabolites are identified correctly in the first position (magenta

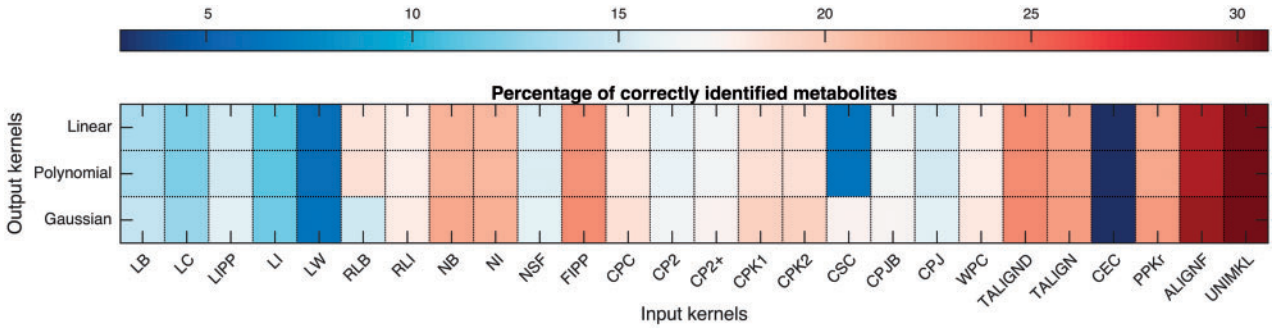

Fig. 4. Heatmap of the percentage of correctly identified metabolites (Top 1) with IOKR. The rows correspond to the different output kernels built on fingerprints (linear, polynomial and Gaussian) and the columns to the 24 input kernels derived from spectra and fragmentation trees, as well as the two multiple kernel combination schemes ALIGNF and UNIMKL

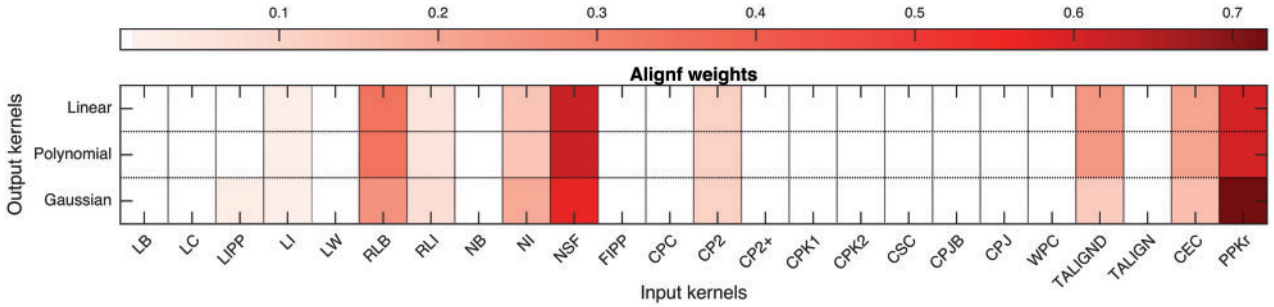

Fig. 5. Heatmap of kernel weights learned by ALIGNF for all pairs of input and output kernels on GNPS dataset. The weights have been averaged over the 10 CV folds

bars) and 71.7% are within the top 10 (cyan bars). The sizes of the candidate sets do not seem to have a strong influence on the identification accuracy. Even for large candidate sets our method is able to identify significant proportions of molecules within top 1 and top 10.

We found 1203 compounds in the GNPS dataset that can be linked to the ontological classification database ChEBI (Hastings *et al.*, 2013). We are interested in evaluating whether there are some classes of compounds we can identify very well and some for which we cannot. Due to the hierarchical nature of the ontological classification, the classes far away from the root are very specific classes and contain very few compounds while the classes close to the root are very generic classes which contain too many compounds. As a result, we restrict the attention to the classes with shortest paths of length 7 from the root node chemical entity (ChEBI id 24431). For those classes, we count how many compounds in the GNPS dataset belong to them and represent the counts as the size of the points in Figure 7. For each compound, the number of candidates and rank of the correct compound are known, so we plot the median number of candidates associated with the compounds in each class on the x-axis and the proportion of cases for which we have correct compounds with rank $\leq$10 on the y-axis. Notice that we only show the classes containing at least 10 compounds.

From the Figure 7, it is clear that the number of candidates associated with the compounds is not a major factor of the identification results. Many classes with larger number of compounds, as shown with larger points, have around 60% of the cases where the identification lies within top 10. There are some classes we can identify very well like *3-aryl-1-benzopyrans* (ChEBI id: 50753), also called *isoflavonoids*, and *heterocyclic antibiotics* (ChEBI id: 24531), while some classes, shown at the bottom of the figure, contain compounds that are more difficult to identify with our method. Among the difficult cases, there are the compounds belonging to the *cyclic amide* (ChEBI id: 23443) class and to the *cyanides* (ChEBI id: 23424) class. The compounds in the cyanide class contain a cyanid-anion sidegroup, which corresponds to a carbon atom connected to a nitrogen atom via a triple bond.

We also studied the differences in prediction performance between CSI:FingerID and IOKR for the different compound classes. A detailed plot showing the differences between the numbers of compounds better ranked by the two methods is given in the Supplementary Materials. This plot shows that IOKR obtains better performances than CSI:FingerID in 74% of the classes. Interestingly IOKR presents the highest improvement for the *cyanides* class and one of its child. On the opposite CSI-FingerID considerably improves the performance for the compounds belonging to the *heterocyclic antibiotics* class and two of its children.

4 Discussion

In this article, we have proposed for the first time to solve the metabolite identification problem using a structured output prediction method, namely IOKR. We have shown that our method improves the metabolite identification rate comparing to competing methods with considerable shorter running time, in practise allowing training the models on a single computer instead of a large computing cluster. In addition, the structured output approach provides a more streamlined—and thus more easy to maintain—one-step prediction pipeline, as opposed to two-step pipelines of CSI:FingerID and FingerID which call for predicting and scoring fingerprints as an intermediate step.

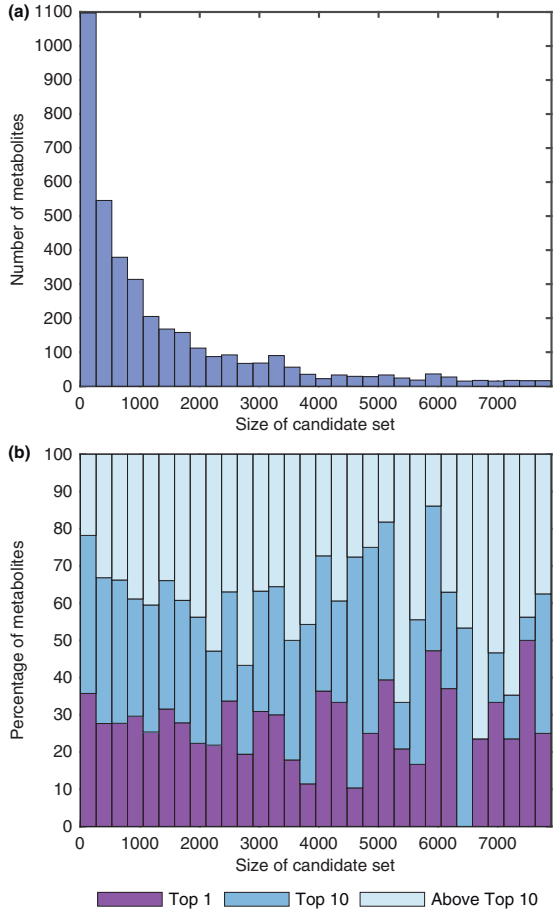

Fig. 6. Identified metabolites with IOKR in function of the size of candidate sets. We considered the candidate sets of size smaller than 8000, which corresponds to 98.8% of the sets, and divided them in 30 bins according to their sizes. **(a)** indicates the number of test metabolites that have a candidate set size in the corresponding size bin. The percentage of metabolites that are ranked in top 1 position, top 10 or above is shown on the **(b)** for the test metabolites falling in each size bin

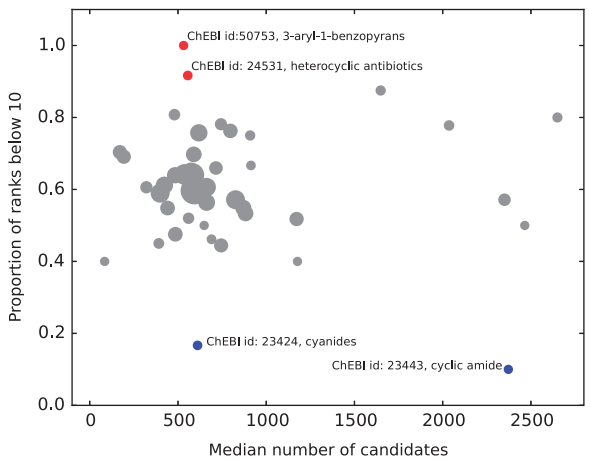

Fig. 7. Scatter plot of classes in ChEBI ontology with shortest paths of length 7 from the class chemical entity. X-axis corresponds to the median number of candidates associated with the compounds in each class and y-axis to the proportion of correct compounds with rank less or equal to 10 for each class. The size of the point is proportional to the number of compounds in GNPS dataset that belong to that class and we only show classes with at least 10 compounds. The classes we can identify well are shown in red and the classes we cannot are shown in blue with ChEBI id and name next to them

For future work, the most important direction is to address the prediction of the 'dark matter' in metabolomics (da Silva *et al.*, 2015): the metabolites that fall outside the compounds in molecular structure databases. There, we need to design better kernels and preimage algorithms for molecular structures.

Finally, it is important to note that the recent breakthroughs in machine learning methodologies for metabolite identification rely heavily on the existence off community efforts building open reference databases such as GNPS and Massbank. At the same time, the reference databases still cover a small fraction of relevant metabolite space. Although machine learning can generalize and extrapolate beyond the training data, as also shown in this article, the scarceness of training data still imposes limits on how accurate models can be built. To really push metabolomics forward, we should widen and make more systematic the community efforts in building and utilizing reference databases.

Acknowledgement

We acknowledge the computational resources provided by the Aalto Science-IT project.

Funding

This work has been supported by the Academy of Finland [Grant 268874/MIDAS] (to C.B., H.S. and J.R.) and the Deutsche Forschungsgemeinschaft [Grant BO 1910/16] (to K.D.).

Conflict of Interest: none declared.

References

Allen,F. *et al.* (2014) CFM-ID: a web server for annotation, spectrum prediction and metabolite identification from tandem mass spectra. *Nucleic Acids Res.*, **42**, W94–W99.

Allen,F. *et al.* (2015) Competitive fragmentation modeling of ESI-MS/MS spectra for putative metabolite identification. *Metabolomics*, **11**, 98–110.

Böcker,S. and Rasche,F. (2008) Towards de novo identification of metabolites by analyzing tandem mass spectra. *Bioinfomatics*, **24**, i49–i55.

Bolton,E. *et al.* (2008). PubChem: Integrated platform of small molecules and biological activities. *Chapter 12 in Annual Reports in Computational Chemistry*, vol. **4**, pp. 217–241.

Brouard,C. *et al.* (2011). Semi-supervised penalized output kernel regression for link prediction. In *Proceedings of the 28th International Conference on Machine Learning*, pp. 593–600. Bellevue, Washington, USA.

Brouard,C. *et al.* (2015). Input output kernel regression: supervised and semi-supervised structured output prediction with operator-valued kernels. Technical Report hal-01216708.

Cortes,C. *et al.* (2005). A general regression technique for learning transductions. In *Proceedings of the 22nd International Conference on Machine Learning*, pp. 153–160. ACM, New York, NY, USA.

Cortes,C. *et al.* (2012) Algorithms for learning kernels based on centered alignment. *J. Mach. Learn. Res.*, **13**, 795–828.

da Silva,R.R. *et al.* (2015) Illuminating the dark matter in metabolomics. *Proc. Natl. Acad. Sci. USA*, **112**, 12549–12550.

Dührkop,K. *et al.* (2015) Searching molecular structure databases with tandem mass spectra using CSI:FingerID. *Proc. Natl. Acad. Sci. USA*, **112**, 12580–12585.

Evgeniou,T. *et al.* (2005) Learning multiple tasks with kernel methods. *J. Mach. Learn. Res.*, **6**, 615–637.

Geurts,P. *et al.* (2006) Kernelizing the output of tree-based methods. In *Proceedings of the 23th International Conference on Machine learning*, pp. 345–352. Pittsburgh, Pennsylvania, USA.

Hastings,J. *et al*. (2013) The ChEBI reference database and ontology for biologically relevant chemistry: enhancements for 2013. *Nucleic Acids Res*., **41**, D456–D463.

Heinonen,M. *et al*. (2008) FiD: A software for ab initio structural identification of product ions from tandem mass spectrometric data. *Rapid Commun. Mass Spectrom*., **22**, 3043–3052.

Heinonen,M. *et al*. (2012) Metabolite identification and molecular fingerprint prediction through machine learning. *Bioinformatics*, **28**, 2333–2341.

Hill,A.W. and Mortishire-Smith,R.J. (2005) Automated assignment of high-resolution collisionally activated dissociation mass spectra using a systematic bond disconnection approach. *Rapid Commun. Mass Spectrom*., **19**, 3111–3118.

Horai,H. *et al*. (2010) MassBank: A public repository for sharing mass spectral data for life sciences. *J. Mass Spectrom*., **45**, 703–714.

Kadri,H. *et al*. (2010) Nonlinear functional regression: a functional RKHS approach. In *Proceedings of International Conference on Artificial Intelligence and Statistics*, pp. 111–125. Sardinia, Italy.

Kadri,H. *et al*. (2013) A generalized kernel approach to structured output learning. In *International Conference on Machine Learning (ICML)*, pp. 471–479. Atlanta, USA.

Lim,N. *et al*. (2014) Operator-valued kernel-based vector autoregressive models for network inference. *Mach. Learn*., **99**, 489–513.

Lindsay, R. *et al*. (1980) *Applications of Artificial Intelligence for Organic Chemistry: The DENDRAL Project*. McGraw-Hill, New York.

Marchand,M. *et al*. (2014) Multilabel structured output learning with random spanning trees of max-margin markov networks. In *Advances in Neural Information Processing Systems*, pp. 873–881. Montreal, Canada.

Micchelli,C.A. and Pontil,M.A. (2005) On learning vector-valued functions. *Neural Comput*., **17**, 177–204.

Ridder,L. *et al*. (2013) Automatic chemical structure annotation of an LC–MS*n* based metabolic profile from green tea. *Anal. Chem*., **85**, 6033–6040.

Rousu, J. *et al*. (2007) Efficient algorithms for max-margin structured classification. In *Predicting Structured Data*, pp. 105–129. MIT Press.

Senkene,E. and Tempel'man,A. (1973) Hilbert spaces of operator-valued functions. *Lithuanian Math. J*., **13**, 665–670.

Shen,H. *et al*. (2013) Metabolite identification through machine learning–tackling CASMI challenge using FingerID. *Metabolites*, **3**, 484–505.

Shen,H. *et al*. (2014) Metabolite identification through multiple kernel learning on fragmentation trees. *Bioinformatics*, **30**, i157–i164.

Su,H. and Rousu,J. (2015) Multilabel classification through random graph ensembles. *Mach. Learn*., **99**, 231–256.

Taskar,B. *et al*. (2004) Max-margin Markov networks. In *Advances in Neural Information Processing Systems (NIPS)*, vol. **16**, p. 25. Vancouver, Canada.

Tsochantaridis,I. *et al*. (2004) Support vector machine learning for interdependent and structured output spaces. In *International Conference on Machine Learning (ICML)*, p. 104. Banff, Canada.

Wang,Y. *et al*. (2014) MIDAS: a database-searching algorithm for metabolite identification in metabolomics. *Anal. Chem*., **86**, 9496–9503.

Weston,J. *et al*. (2003) Kernel dependency estimation. In Becker, S. Thrun, S., and Obermayer, K. (eds.) *Advances in Neural Information Processing Systems 15*. MIT Press.

Wolf, S. *et al*. (2010) In silico fragmentation for computer assisted identification of metabolite mass spectra. *BMC Bioinformatics*, **11**, 148.