The unresolved problems of peptide drug metabolism. Software-aided approach designed to analyze and predict cleavage sites for natural and synthetic peptides.

Tatiana L. Radchenko

DOCTORAL THESIS UPF / 2018

THESIS DIRECTORS

Dr. Ismael Zamora Rico (Lead Molecular Design S.L.);

Dr. Manuel Pastor Maeso

DEPARTAMENT CEXS

upf. Universitat Pompeu Fabra Barcelona

To my lovely Vadim and beloved family

Acknowledgments

I was very fortunate to perform my thesis work at a company as collaborative as Lead Molecular Design S.L.; and, there are many people to thank for their part in my success. I would first like to thank my supervisor, Dr. Ismael Zamora Rico, for giving me home in his company and support over the years. I would like to thank him for believing in me at our first meeting and giving an opportunity to start this project and go on with it. Under his mentorship I have learned of scientific ideas generation and realization in life without fear of failure, which is an invaluable tool for the whole live forward. I would like to acknowledge his readiness to support all my intents in self-development: additional courses and reading extra books. I would also like to thank my thesis director, Dr. Manuel Pastor Maeso for his contributions to this work. Over the years, he has given me scientific guidance and suggestions. He demonstrated a sincere interest in my work. I am fortunate to work in such a group of intelligent scientists and software developers from Lead Molecular Design S.L. who helped me in all areas of the thesis work – scientific consulting, help in programming, software design and algorithmic solutions. I would like to thank you all of you – Fabien Fontaine, Luca Morettoni, Blanca Serra, Xavi Pascual, Elisabeth Ortega, Guillem Plasencia, Vera Lopez, Esther Gilabert. Blanca and Elisabeth have always been there to talk if ever something was bothering me regarding my work or in life. Fabien and Luca have always been there to help with coding and compilation problems. Also, I would like to thank Fabien for his contributions to all articles included in this thesis. I would like to recognize the members of the Molecular Discovery who have all contributed to this thesis work. Especially Gabriele Cruciani for his support and resources provided to complete needed experimental and analytical work. I would like to give a special thank you to Aurora Valeri, Laura Goracci and Lydia Siragusa for their incredible efforts and help with experimental work and sharing their knowledge on other topics. The Pharmacokinetics, Pharmacodynamics, and Drug Metabolism, Merck & Co., Inc lab members wonderful group of people who I have enjoyed working with and learning from. Especially I would like to thank Christopher Kochansky who taught me several techniques and aided in experiments, who shared his enthusiasm for science with me, who helped me a lot by correcting several papers submitted during this thesis. I would like to thank the Pharma Research and Early Development, Roche Innovation Center Basel, F. Hoffmann-La Roche Ltd members, especially Andreas Brinks, Yves Siegrist, who have given their time to teach me several techniques and helped in experimental work. Alison Bateman deserves huge thanks for dealing with experimental work. I would like to acknowledge Kevin Bateman for his support and mentorship during this project and his fantastic hospitality during my stay in West Point. I would like to acknowledge the Russell J. Mortishire-Smith and Jayne Kirk for their help with Ion Mobility project and always being excited to talk about scientific problems or data. I would like to express my appreciation for the optimistic attitude that Rosa Martinez and Lidia Mateo brought each time we met and discussed our projects. I would like to acknowledge the support from Doctorats Industrials, AGAUR, Generalitat de Catalunya.

I would like to express the deepest gratitude to my family. My mother Natalia, father Lev, sister Anna and brother Grigorii - you have all provided support, encouragement and interest in my work. I would not be who am I today without you all. Finally, I would like to thank my husband, Vadim. You have been continually supportive of my whole change of my carrier path. Thank you for teaching me to avoid thinking on “If” and to stay positive, for yours advises on the algorithm development part, for listening, for patience and for making all this tea.

Abstract (English)

The work of this thesis aims to develop of a new workflow to predict potential cleavage sites in new candidate peptide drugs. The main goal of this workflow is to understand the protease specificity rules from data coming from different sources (experimental data and/or external databases) and with no limitations on peptides structure (linear/cyclic, containing natural and/or unnatural amino acids) or specific experimental conditions (individual proteases or complex matrices such as plasma).

At the first step we implemented a new algorithm to store the information from experimental and external sources in a chemically aware database WebMetabase, so that the chemically aware exact substructure and a similarity-based substructure search can be performed. The main advantages of this database are that it allows to combine data from different sources and can be further enriched with new data without limitations. Moreover, because each peptide structure was interpreted as a chemical structure, we were able to process information about cyclic peptides. Since each amino acid was described by pharmacophoric and physicochemical properties, there were no limitations for the processing of the unnatural amino acids.

At the second step we developed and applied a frequency analysis approach to reveal the metabolically labile amide bonds defined by similar pharmacophoric or physicochemical properties of residues around the cleavage site towards the specific proteases/specific media.

At the third step we built several predictive models using a training dataset, where each site of cleavage was described by its molecular descriptors and frequency. Therefore, prediction ability of the models was not limited to only natural amino acids in contrast to the state-of-the-art approach such as PeptideCutter. We demonstrated that models can be trained for different proteases on MEROPS exported data and have a comparable predictive performance as other public tools such as PROSPERous.

Resumen (Spanish)

El trabajo realizado durante la presente tesis doctoral se centra en el desarrollo de un sistema para la predicción del sitio de catabolismo de péptidos. El objetivo fundamental del mismo es el encontrar patrones de especificad para el sitio de metabolismo producidos por proteasas considerando diversas fuentes de datos (desarrolladas experimentalmente internamente por el usuario o publicadas en base de datos externas), sin limitaciones en la estructura de los péptidos (lineal/cíclicos, con aminoácidos naturales o no) y que sean capaces de considar varias condiciones experimentales (incubaciones con proteasas individualmente o en matrices complejas como plasma)

En un primer paso, se implementó un nuevo algoritmo para poder guardar la información de manera sistemática independientemente de la fuente en una base de datos que considera la estructura química, WebMetabase. De esta manera se pueden realizar búsquedas sub-estructurales o de basadas en similitud sub-estructural. La principal ventaja de la estrategia seguida es que permite combinar datos de diversas fuentes y por lo tanto puede ser actualizada con nuevos datos sin ningún tipo de limitación. Además, al guardar la estructura de los péptidos como estructuras químicas y no tan solo como secuencias de monómeros, es posible procesar compuestos cíclicos de cualquier tipo. En este procedimiento cada aminoácido se caracteriza mediante una serie de propiedades fisicoquímicas y farmacofóricas, de tal manera que no existe limitaciones a la hora de comparar aquellos monómeros de procedencia natural o de síntesis.

En un segundo paso, se ha desarrollado y aplicado a la base de datos antes mencionada un análisis de frecuencias que permite definir las propiedades fisicoquímicas y/o farmacofóricas de los residuos que participan en los enlaces amíSdicos metabólicamente lábiles sobre cualquier proteasa o medio de incubación.

Por último, se han construido modelos predictivos que proporcionan la probabilidad (frecuencia) de que un enlace sea objeto de reacciones metabólicas de hidrólisis. Estos modelos se han derivado a partir de conjuntos de aminoácidos caracterizados con los descriptores moleculares ya mencionados y que tienen asociados la frecuencia obtenida del análisis anterior. De esta manera, la

capacidad predictiva no se encuentra limitada a la estructura de los aminoácidos naturales como ocurre con alguno de los programas que se consideran más avanzados en la actualidad, como, por ejemplo, PeptideCutter. Múltiples modelos se han obtenidos a partir de conjuntos de datos exportados de la base de datos MEROPS. Estos modelos presentan una capacidad predictiva comparable con herramientas de acceso público como por ejemplo PROSPERous, pero sin sus limitaciones en cuanto a fuente de información ni estructura de los péptidos tratados.

Preface & Justification

Peptides therapeutics are becoming increasingly important on the pharmaceutical market. However, it is known that peptide drugs bioavailability and stability are lower than for small molecules. It is highly important to understand peptide drug metabolism and optimize its clearance as it influences the drug safety and efficacy. Because peptides are mainly cleaved by peptidases, every new candidate must be designed considering the localization of potential protease sites of cleavage. The information about proteases and their sites of cleavage is widely spread across publications and databases. Although useful, these databases still have several limitations. For example, none of the available resources allow to add new information in an automatic way to enrich the database.

In this study, we developed an approach based on Mass-MetaSite and WebMetabase to process data-dependent and data-independent acquisition high-resolution mass spectrometry data from *in vitro* incubation samples, to elucidate metabolites structures, to predict cleavage sites and to store the results in a chemically aware database. Furthermore, we added a new method that processes the information from external sources. After processing received data on peptide substrate and metabolites is annotated in accordance with new developed annotation system and persisted in WebMetabase. The annotation of the peptide information in this manner enables a chemically aware exact substructure search and a similarity-based substructure search inside the database. Moreover, we implemented an algorithm that performs frequency analysis and similarity frequency analysis approach that reveals the group of metabolically labile amide bonds defined by exact structure matching or similar molecular properties towards the specific proteases.

In this study analyzing several datasets of data for cyclic and linear peptides, containing nonstandard amino acids collected from different sources we demonstrated how frequency analysis could be used to build predictive models and how these models can be applied to predict the metabolic liability of different amide bonds in a new non-tested peptide. These results were described in the following publications:

1. “Software-aided approach to investigate peptide structure and metabolic susceptibility of amide bonds in peptide drugs based on high resolution mass spectrometry”

2. “Software assisted analysis for peptide drug metabolism”

3. “WebMetabase: cleavage sites analysis tool for natural and unnatural substrates from diverse data source”

4. “Software-aided workflow for predicting protease-specific cleavage sites using physicochemical properties of the natural and unnatural amino acids in peptide-based drug discovery.”

5. “Metabolite Identification Using A Ion-Mobility Enhanced Data Independent Acquisition Strategy and Automated Data Processing.”

Therefore, one of the main advantages of this approach is that it generates a searchable database for the information coming from different sources that can be enriched with new data. Nevertheless, the proposed methodology as opposed to existing databases can be applied in the case of non-natural amino acids and/or cyclic peptides. Moreover, since the system used to derive the cleavage site appearance rules could be enriched with the new experiments, models can be re-trained with updated dataset and the derived rules can be refined to tune the system for the experimental conditions and/or peptide families of interest. This knowledge can be applied during the design-make-test drug discovery cycle.

Table of contents

OBJECTIVES

General objective:

To obtain and validate a new workflow to generate protease specificity rules and use them to predict potential cleavage sites in new candidate peptide drugs. The proposed metholodology should work without any limitations on: peptide structure (cyclic or linear, containing natural and unnatural amino acids); experimental procedure (specific protease or complex matrix); analytical mass spectrometry acquisition technique (data-dependent or data-independent methods); and source of data (experimental or external).

Specific objectives:

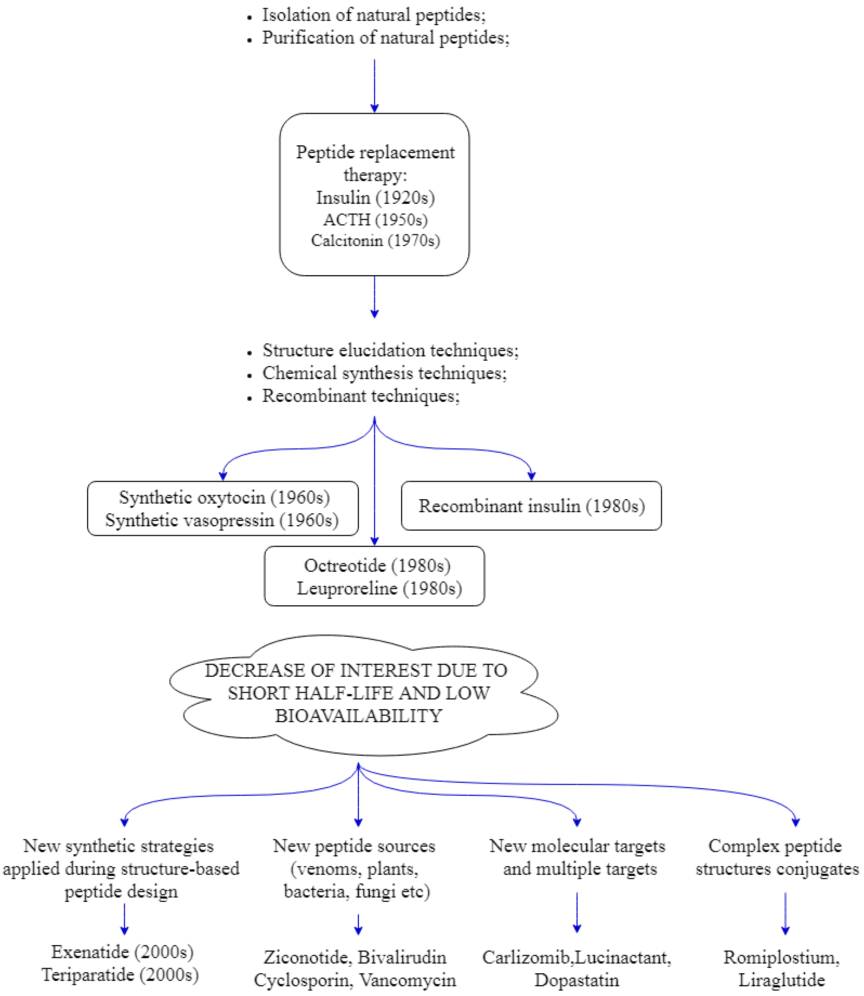
1. Develop an approach to process data from different incubation samples for any kind of peptide structure and to perform metabolite identification for processed samples for multiple acquisitions MS data.
2. Develop an approach to annotate processed information from multiple sources, suitable for interpreting data on any type of amino acid and to store annotated data into the searchable database.
3. Implement a frequency analysis algorithm which, using the created annotation system is able to reveal the protease specificity rules.
4. Explore the usage of the described system for the development of site of cleavage predictive models useful for peptide drug design specialists.

INTRODUCTION

1.1 Peptides as therapeutics

During the last two decades, the interest in peptide therapeutics has increased in pharmaceutical research and development. Peptides position in the pharmaceuticals molecule space is between small molecules and proteins. Peptides are defined as polypeptides containing from 2 to 50 amino acids (aa) but differ biochemically and therapeutically from small molecules and proteins [1,2]. Peptides can act naturally as hormones, neurotransmitters, growth factors or antibacterial agents. They are generally thought to be well-suited for diseases where the target is a protein-protein interaction [3, 4]. Peptides have great potential as new drugs due to a good safety and tolerability profile, a higher efficacy and selectivity comparing to small molecules [1-3,5-7] and high specificity to certain protein targets, for example G-protein-coupled receptors (GPCR) [1,5,6].

Recently, peptides became an important element on the pharmaceutical market. Peptide therapeutics timeline on pharmaceutical market is shown in Figure 1. The initial appreciation of the peptide drugs was related to the fact that native peptides could be used as a replacement therapy in case of lack or absence of endogenous hormone. The first peptide administered as a drug was insulin extracted from animal pancreas in the 1920s. Insulin became the first commercially available peptide therapeutic [8]. Since then several more natural peptides such as adrenocorticotropic hormone (ACTH) and calcitonin have been used as a hormone therapy [9]. Later when synthetic strategies became available synthetic oxytocin, vasopressin and octreotide (synthetic analogue of somatostatin) entered in the pharmaceutical market. In 1982 when the recombinant technique was developed the first recombinant insulin was introduced [8]. In 1988 one of the most important cell permeable peptides (CPP) - trans-activator of transcription protein (TAT) from human immunodeficiency virus-1 (HIV-1) was discovered [10]. Moreover, venoms, plants, bacteria and fungi were considered as a new source for isolation of new natural peptide therapeutics.



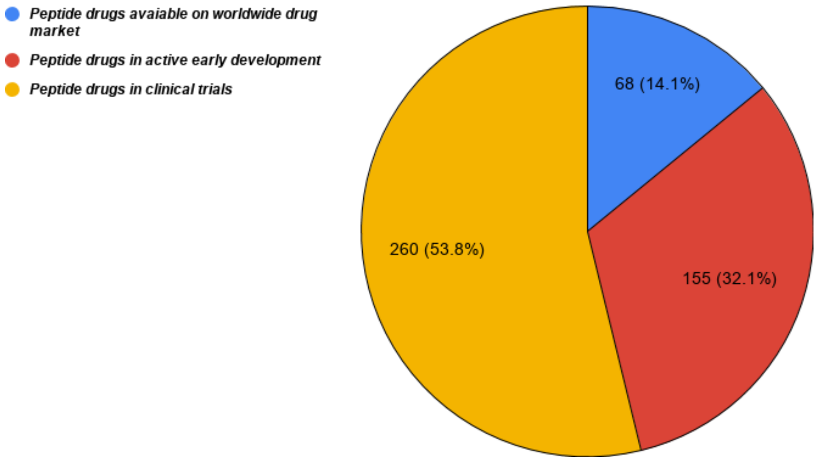
Isolation of natural peptides;
Purification of natural peptides;
Peptide replacement therapy:
Insulin (1920s)
ACTH (1950s)
Calcitonin (1970s)
Structure elucidation techniques;
Chemical synthesis techniques;
Recombinant techniques;
Synthetic oxytocin (1960s)
Synthetic vasopressin (1960s)
Recombinant insulin (1980s)
Octreotide (1980s)
Leuproreline (1980s)
DECREASE OF INTEREST DUE TO SHORT HALF-LIFE AND LOW BIOAVAILABILITY
New synthetic strategies applied during structure-based peptide design
New peptide sources (venoms, plants, bacteria, fungi etc)
New molecular targets and multiple targets
Complex peptide structures conjugates
Exenatide (2000s)
Teriparatide (2000s)
Ziconotide, Bivalirudin
Cyclosporin, Vancomycin
Carlizomib,Lucinactant,
Dopastatin
Romiplostium,
Liraglutide

Figure 1. Peptides utilization timeline in drug development

The new knowledge in genomics and proteomics provided additional information about potential peptide receptors targets [9]. In addition, other more complex peptide structures such as conjugates gained popularity and entered in clinical development since 2010. Conjugation was used as a methodology to increase half-life time and improve stability profile of the peptide drugs. From a chemical point of view, peptide drugs can be divided in three main groups (native, analogues and heterogenous), where heterogenous are the peptides discovered through synthetic library screening or phage display [9] or using other methods but independently from the native peptides.

Today, 68 peptides are represented on the worldwide drug market, approximately 155 peptide-based drugs are in active early development and 260 are tested in clinical trials [1, 5, 9]. It is worth mentioning that about 28 peptide drugs were approved worldwide since 2000s [8] (Figure 2). Recently the most successful peptide drugs available on the market are Copaxone, Victoza, Lupron, Zoladex, and Sandostatin [2, 11]. Peptide drugs are well-suited for treatment in a wide range of therapeutic areas, such as diabetics, cancer, osteoporosis, hormone therapy, cardiovascular diseases, anemia, bowel syndrome, Cushing’s disease, multiple sclerosis, HIV, and many more [3].

Figure 2. Development status of therapeutic peptides.



Depending on peptide drug target localization and/or amount of the targets peptide therapeutics can be classified in three main groups:

- Peptide drug with extracellular target (most of the peptide drugs available on the market)
- Intracellular peptides (CPPs used for drug delivery)
- Hybrid peptides (Dopastatin, “Twincretins”)

The largest group of peptide drug targets is represented by GPCRs, the remaining targets are ion channels, and other extracellular targets such as structural proteins or secreted enzymes. Big portion of peptides available on the market or in development is represented by antimicrobial peptides. Only few peptides, so called cell penetrating peptides, are used to treat intracellular targets in drug delivery [9, 12, 13]. CPPs are short amino acid oligomers (5-30 residues) that contain or correspond to protein transduction domains (PTDs). The most important feature of CPPs is their ability to directly permeate the cell membrane and consequently the ability to deliver substances (e.g. proteins, antibodies, small molecule drugs etc.) into cells [12, 13]. As any peptide, CPPs have advantages comparing to small molecules, such as low cytotoxicity even at the concentration at which CPP permeation occurs and are metabolically degraded after delivery of the accompanied substance. Also, several peptides were developed as multitarget drugs. For example, Dopastatin was developed to address two targets involved in neuroendocrine tumor disease pathology and was synthetized as a hybrid of peptidic somatostatin receptor agonist linked to a small molecule dopamine agonist [14]. Other hybrid molecules are “Twincretins”, they are glucagon-like peptide-1 (GLP-1) agonist and the glucagon of gastric inhibitory polypeptide (GIP) receptors agonist [15]. Another peptide that acts as GLP-1 agonist covalently bonded to a proprotein convertase is subtilisin/kexin type 9 (PCSK9)-inhibiting antibody [16].

1.1.1 Peptide drug advantages and disadvantages

Since many native peptides are natural ligands for many cell surface receptors, they are acting as agonists in many pathways and therefore recognized as a potent and highly selective candidate drug. Currently, most of the peptide drugs are represented by native peptide analogues since their absorption, distribution, metabolism and elimination (ADME) properties, safety and toxicity profiles are known and easier to predict. On the contrary, synthetic peptides require more attention to the analysis of ADME and toxicity (ADMET) properties since the chemical structures of some of the canonical monomers used for the synthesis are modified and it is necessary to evaluate the potential ADMET properties of the new molecule produced.

Limited development of therapeutic peptides occurred in the past due to insufficient ADME properties: low permeability, low solubility, short half-life time and limited residence time in tissues. Low cell permeability is often related to structural factors such as high hydrogen bonding capacity and low lipophilicity [3]. Because of this limitation the therapeutic application of the peptide drugs is restricted to extracellular and transmembrane agents, excluding the group of peptides so-called cell penetrating peptides. Another permeability issue is that peptides cannot easily pass the gastrointestinal (GI) barrier, thus parenteral administration route is a preferable comparing to oral and that one is not the most convenient and compliant route of administration. Moreover, peptides cannot cross blood-brain barrier (BBB) and it excludes central nervous system (CNS) drug targets [11]. Therefore, new techniques were developed and applied to improve permeability properties of the peptides described below. Moreover, low oral bioavailability is more frequently related to physiological processes. Since natural peptides exist as a part of the natural pathways they act as agents in well-established system. It means that they are produced as an answer to the biological signal, processed, released, they perform their function and then are rapidly metabolized to turn off the signal. Fast extraction happens through proteolysis, and pH dependent hydrolysis in blood, GI tract, and liver with consequent renal filtration [1,3,7,17]. Therefore, natural peptides have in general a short half-life. On one hand it is a limitation because peptides cannot be administered orally and should be administered through injection or delivered via non-oral routes such as trans buccal, nasal, inhaled or transdermal [1-3,5,7,17]. On the other hand, it means that drug-drug interactions are rarely observed, and the toxicology profiles seem to be safer than for small molecules because peptides do not accumulate in the tissue and metabolism in liver is not generally significant. [8].

In case that the peptide drug is administered orally there are several enzymatic barriers that should be crossed to become a successful drug (Figure 3).

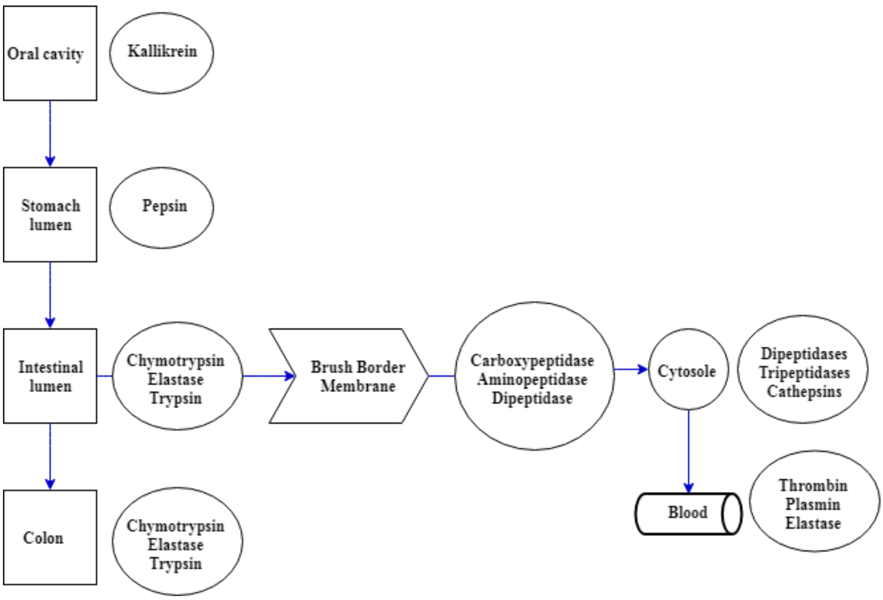


Figure 3. Enzymatic barriers that should be considered during drug development of orally administered peptide

It is well known that numerous human proteases are involved in peptide degradation. The most important barrier after oral administration is the lumen of the small intestine, which contains peptidases secreted from the pancreas (e.g. chymotrypsin), as well as cellular peptidases from mucosal cells. The second one would be the brush border membrane of the epithelial cells, which contains at least 15 different peptidases [18]. Therefore, several structure-based peptide design methodologies described below were developed to improve the stability of the peptide drug.

1.1.2 Peptide drug structure-based design steps

During rational peptide therapeutics design the following traditional structure-based design steps are usually completed (Figure 4). These steps are essentially important for identifying possible sites of substitution since these chemically labile residues can be isomerized, glycosylated or oxidated:

- identification of minimum active sequence (e.g. known crystal structure of the peptide with given secondary and tertiary structure) and selection of the lead compound;

- positional scanning to determine critical amino acids through alanine substitution (Ala-scan) and analysis of small focused libraries and structure-activity relations (SAR),

- protection from degradation at the terminal ends with modification of the C- and N-terminal of the peptide preventing degradation by carboxy- and aminopeptidases,

- identification of the sites of cleavage (e.g. during metabolites elucidation) and finally chemical modifications applied during structure-based peptide design. These modifications are implemented to improve the following peptide candidate drug properties: selectivity, solubility, stability, bioavailability, safety and toxicity [11].

- And finally, perform substitution of amino acids and building a SAR via experiments such as an alanine scan and estimation the half maximal effective concentration (EC50) for each of the modified compounds.

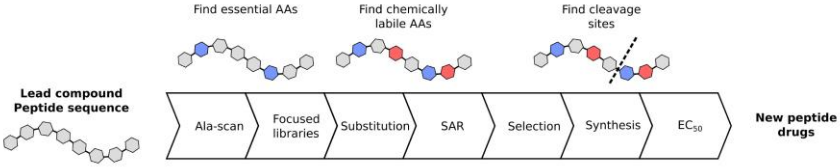


Figure 4. The traditional structure-based design strategies that are used in peptide drug discovery.

Currently, several strategies are applied to select peptide lead compounds:

- Native peptides used as a starting point;

- Peptides derivation from natural products;

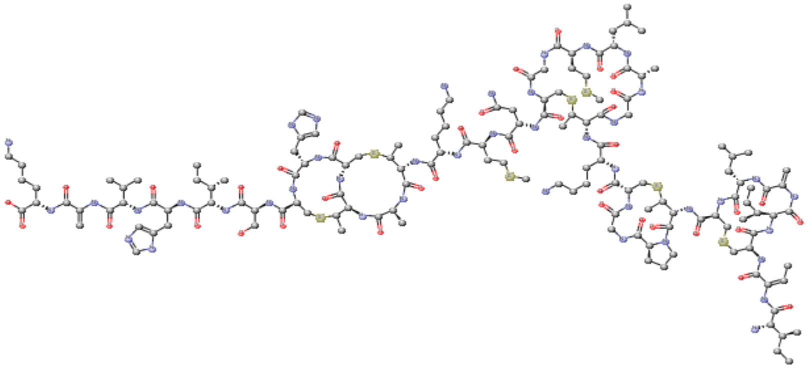
- De-novo peptide discovery;

- Peptide discovery through genomic, proteomic, peptidomic approaches.

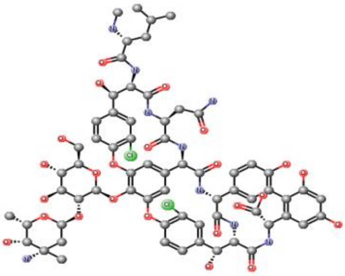
The first and most common is when a native known peptide molecule is used as a starting point. Native peptides are synthetized by ribosome and composed from canonical standard amino acids connected through peptide bonds. It can be a peptide with a known target of interest such as hormone. In this case the main efforts are spent on optimizing its ADME properties, increasing the half-life and/or improving the stability and selectivity using synthetic chemical modifications. Octreotide and leuprolide were discovered as synthetic analogues of the natural hormones somatostatin and luteinizing-hormone releasing hormone (LHRH), respectively. The second strategy would be a derivation of a peptide candidate drug from natural product and its optimization to improve its ADME properties [11,19,20]. The third strategy would be de-novo peptide discovery. In this case the screening is performed on synthetically or biologically (phage, mRNA) produced compound libraries [10-13]. The fourth and last strategy is using genomic, proteomic, peptidomic approaches. In this case the efforts are spent in finding peptides of interest in different species or finding a new activity for known peptides [11,21].

Natural product derived peptides provided productive starting points for many drug discovery strategies since they contain structures that aim to improve peptide membrane permeability properties, stability and bioavailability comparing to the native ones. For example, non-ribosomal synthesized peptides controlled by non-ribosomal peptide synthetases (e.g. vancomycin, structure is presented in Table 1) produced by bacteria and fungi can contain nonstandard amino acids (e.g. N- and D-methylated aa) and cyclization motifs [11,19,20]. Peptides extracted from venoms are usually cyclic disulfide-rich sequences. Several optimized peptides reached the market such as captopril, exenatide and ziconotide [11,22-24]. Ziconotide structure is shown in Table 1. Plant-derived cyclotide peptides are characterized as thermally stable and orally active peptides. They are organized as a cyclic head-to-tail structure with disulfide bridges and demonstrate enhanced stability and bioavailability in plasma and gastric fluids. Lanthipeptides such as nisin are cyclic peptides that contain macrocyclic thioether linkage [11,25-27]. Nisin structure is shown in Table 1. All these structural patterns and motifs were used in peptide medicinal chemistry optimization strategies and structure-based peptide design methodologies.

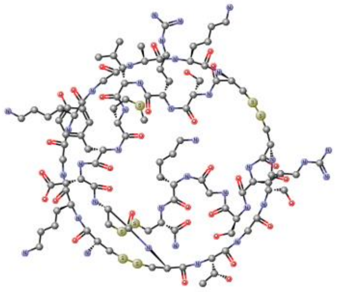
Table 1. Structures of natural peptides



Nisin



Vancomycin



Ziconotide

1.1.2.1 Peptide drug optimization techniques

To achieve better ADME properties, improve solubility, reduce aggregation tendency the following chemical modifications are typically applied: substitution of the common L-amino acids to D-amino acids or other unnatural amino acids, backbone N-methylation, alpha-methylation of amino acids, incorporating of beta-amino acids and others [3,7,11,28,29] (Table 2).

Since hydrogen-bonding capacity (especially intramolecular hydrogen bonding), hydrophobicity/lipophilicity, size, and polar surface significantly influence peptide permeability, it was found that the introduction of hydrogen bond acceptor–donor pairs in peptides can improve membrane permeability [30] (Table 2).

To maximize half-life time of peptide drug the usual modification technique is implemented to introduce a limitation in the enzymatic degradation of the peptide. For this purpose, the possible cleavage sites should be identified, and it should be followed by substitution of identified residues and/or protection against proteolytic degradation through enhancement on the secondary structure (e.g. insertion of a structure inducing probe (SIP)-tail, lactam bridges, stapling or clipping of peptide sequences or cyclization) [1,16] (Table 2). Several strategies were developed to specifically increase half-life of peptide in plasma such as peptide acylation, insertion of albumin-binding peptide elements in the peptide backbone, conjugation to albumin-binding antibody fragments [1] (Table 2). Also, the N-terminus residue of a peptide correlates to its half-life in plasma. For example, if peptides contain Met, Ser, Ala, Thr, Val, or Gly as N-terminus, they have longer half-lives. On the contrary, half-life is shorter if peptides contain Phe, Leu, Asp, Lys, or Arg as N-terminus. Moreover, peptide sequences rich in Pro, Glu, Ser, and Thr are more sensitive to enzymatic degradation [3]. To reduce renal elimination, polyethylene glycol (PEG)-ylation can be used to reduce globular filtration [35].

Table 2. Strategies applied to improve peptide ADME properties

To improve solubility and reduce aggregation tendency	Chemical modifications:
	Substitution of the common L-amino acids to D-amino acids or other unnatural amino acids;
	Backbone N-methylation;
	Alpha-methylation of amino acids;
	Incorporating of beta-amino acids;
	Salt-bridge formation;
	Cyclization of the peptide;
	Deamination;
	Oxidation;
	Isomerization;
	Peptidomimetic bonds incorporating;
	Usage of peptoids (poly-N-substituted glycines);
	Usage of aza-peptides and others

To improve permeability	Introduction of hydrogen bond acceptor–donor pairs

Table 2. Strategies applied to improve peptide ADME properties	
To maximize half-life time	Enhancement on the secondary structure: Insertion of a structure inducing probe (SIP)-tail; Insertion of lactam bridges; Stapling or clipping of peptide sequences; Cyclization Chemical modifications: Substitution of the common L-amino acids to D-amino acids or other unnatural amino acids; Backbone N-methylation; Alpha-methylation of amino acids; Incorporating of beta-amino acids; Deamination; Oxidation; Isomerization; and others.
To maximimze half-live time in plasma	Peptide acylation; Insertion of albumin-binding peptide elements in the peptide backbone; Conjugation to albumin-binding antibody fragments
To reduce renal elimination	Polyethylene glycol (PEG)-ylation

Nowadays, new technologies are under development to produce semisynthetic organisms with expanded genetic code to produce non-canonical proteins [12,31-34]. These changes are applied during the design-make-test drug discovery cycle, with hopes of improving the physicochemical and pharmacokinetics properties of the compound of interest. Synthetic and modified peptides require more attention to be paid to the analysis of ADMET properties since non-canonical monomers can be incorporated, therefore, it is crucial to evaluate these properties rapidly in early development.

1.2 Examination of peptide ADME properties and stability

In accordance with U.S. Food & Drug Administration drug discovery and drug development process consists of the following steps:

- drug discovery and development;

- preclinical research;
- clinical research;
- FDA review;
- FDA post-market safety monitoring.

During each of these stages drug compound should be examined and tested through different assays to evaluate its ADMET properties and safety profile. Several criteria should be considered during assay selection depending on the stage such as available amount of the compound, tissue where the compound testing should be performed, number of samples to be analyzed. For example, in the early discovery and development stages, preclinical research and clinical research the amount of the investigated compound is limited, and highly sensitive analytical approaches are needed to perform qualitative and quantitative evaluation. On the contrary, there is no such limitation in the last stages of the drug development process.

Early preclinical investigation process can be additionally split in five steps:

- target identification;
- target validation;
- hit finding;
- lead finding;
- lead optimization.

Each of these steps involves certain tasks to be completed. During target identification a target molecule search is performed. This molecule should be related to the disease of interest and involved in the disease development process. At the next step the selected target should be tested to find out if it is therapeutically useful and its effect on the disease of interest. After that hit finding is performed to identify a compound that demonstrates interaction to the selected target. During lead finding ADME properties of the hit compound are evaluated to prove that they are reaching the necessary levels and finally the selected compound is optimized to improve its ADME properties. Due to the significant developments in the automatization of the drug discovery process, data processing and analysis especially at the early stages, it starts to be possible to investigate large number of compounds in short time (high-throughput screening (HTS)), therefore, high number of samples should be analyzed by the multiple bioassays to evaluate the properties of the compound of interest.

Speed, selectivity and sensitivity should be considered before the selection of the bioassay and analytical technique to measure the ADME properties, stability, safety and toxicity profile. Moreover, it is important to consider what information is the most critical for answering the questions of interest. For instance, in absorption studies the main interest is about the amount of parent peptide in serum or plasma after administration, thus a parent drug quantification assay would be required. Alternatively, the mechanisms of the route of absorption (e.g., lymphatic or vascular) or the degree of metabolism can be investigated and in this case analysis of both parent therapeutic and metabolites in lymph or blood should be performed. Measurements of the concentrations for the parent drug in tissues of interest and/or the metabolites can be performed during the distribution studies. The following criteria should be also considered to select the bioassay:

- analyte of interest;
- matrix;
- sensitivity and specificity;
- final goal of the bioassay;
- number of samples needed for the bioassay.

The analyte of interest should be identified to understand the ADME characteristics of the peptide therapeutic. In many situations the analyte of interest is the parent molecule. In some cases, ADME properties of primary metabolites could also be of interest because these molecules can be pharmacologically active. In such a situation, it may be necessary to ensure that the methods used can confirm that all the pharmacologically active sites of the molecule are intact. It may also be of interest to elucidate metabolite structures and measure its amounts to understand the differences in ADME of the active components upon metabolism. This can be achieved by radiolabeling of each component with a separate radioisotope and using tissue sampling to obtain highly quantitative tissue concentrations. In some situations, the identity of the metabolites can be unknown at the time of the design of the ADME studies, and thus careful consideration of the analytical methods and study design is required to ensure that all relevant metabolites can be identified and quantitated.

The study design and choice of analytical method is significantly influenced by the matrix. For example, if distribution and elimination studies are performed in solid tissue use of certain methods (e.g., immunoassays and mass spectrometry) can be more difficult, time-consuming and less sensitive comparing to application for these approaches in liquid matrices such as plasma or serum [36]. Additionally, the commonly used approaches applied for tissue investigation can be classified as destructive and non-destructive. The destructive approach is where the sacrifice of individuals or groups of animals at specified time points is followed by analysis. These methods have the disadvantage of preventing serial assessment of tissue distribution within individual animals and can require relatively large numbers of animals. To overcome these limitations non-destructive methods such as imaging, including PET and optical imaging can be used. As it was described before, large number of samples should be evaluated during HTS and therefore large number of animals should be involved in the investigation, therefore, tissue investigation should be avoided on the early stage due to the regulatory authority’s requirements.

The assay sensitivity is a critical factor since it is highly influenced by other aspects of study design, including matrix and analyte. The assay should be sufficiently sensitive to produce reliable quantitative measures of the desired analyte (s) and to be able to measure compound of interest at pharmacologically relevant concentrations with the necessary accuracy and precision. As mentioned above, during early drug discovery stages available amount of the compound is relatively low, therefore, mass spectrometry and nuclear magnetic resonance spectroscopy are preferable analytical approach to be selected.

Assays used to measure ADME properties of the compound of interest can be performed *in vitro, in vivo* and *in silico*. Depending on the result of the assay they can be split in two main groups:

- Bioassays where actual analytical measurement is done without previous preparation of the sample;

- Bioassays that need a sample preparation step before the actual analytical measurement is done.

In the first group the bioassay itself provides an analytical read out that is used to determine the biological activity of a test substance in a biological fluid being the result information on functional and biological response from the test system [37]. For example, immunoassay that is an analytical procedure that measures the concentration of a test substance in a biological fluid. It uses the principle of specific binding of antigen and antibody, where antigen and antibody can be determined using different methods, for example, labeling (by radiotope or colloidal gold etc.) or other techniques (e.g. agglutination, Western blot etc.). Because of the high sensitivity and specificity, immunoassay was a method of preference for the peptide quantification, but the main disadvantage of the method is that it cannot distinguish active and metabolized peptides. Thus, application of immunoassay for metabolism study is limited [37, 38]. It is used for drug analyte sample measurement and for pharmacokinetic (PK) studies [39].

Another example is the labeling approaches. Labeling techniques are also used for pharmacokinetics and metabolism studies. Peptides can contain certain sequences that can be used as a target for labeling. Different labeling strategies are used directly such as radiolabeling (e.g. halogenation or complexation with metallic radioisotopes) [36]. Fluorescent labeling has become a routine procedure used to evaluate physical properties, biodistribution and pharmacokinetics of the compounds of interest. But labels can influence the physicochemical properties of the molecule and it can lead to undesirable oxidation, deamidation or aggregation [36].

The second group contains bioassays are the ones that needs a sample preparation step before the actual analytical measurement is done. For example, in the case of liquid chromatography (LC), mass spectrometry (MS), nuclear magnetic resonance (NMR) spectroscopy some experimental steps are needed (i.e. dissolution, centrifugation etc.). They are described in detail below.

1.2.1 Bioassays applied on peptides

Multiple *in vivo, in vitro* and *in silico* assays have been developed to address the ADME and stability challenges of peptides.

The combination of *in vitro* data and *in vivo* studies can be used to determine whether low oral bioavailability is related to poor absorption or fast first-pass liver extraction. *In vivo* animal models are frequently used to study peptide bioavailability and fraction absorbed in nonsurgical or portal vein cannulated (PVC) animals. Transporter knockout animals (e.g., PEPT1 knockout mice) are used to understand the contribution of uptake transporters in oral absorption [3].

The bioanalysis of peptides can be challenging due to low sensitivity and selectivity, high nonspecific binding and protein binding, low recovery, carryover, solubility, and stability issues [3]. These challenges can be addressed by: adding protease inhibitors to the collection tubes and putting these tubes on wet ice immediately after blood collection. Adding organic solvents, acids, salts can help to prevent further hydrolysis during sample preparation and analysis. Increasing peptide solubility helps to overcome nonspecific binding [40].

Peptide structures and physicochemical properties can be similar either to small molecules or protein. Thus, strategies used to evaluate the *in vivo* clearance of small molecules can be applied to small and lipophilic peptides (cyclosporine) (e.g., via P450-mediated metabolism) [43]. Peptides similar to proteins are eliminated through proteolysis, renal filtration, catabolism, and endocytosis. Ttherefore, allometric scaling can be an effective tool to predict human PK parameters (e.g., volume of distribution and clearance) from preclinical species [3].

1.2.1.1 Proteolytic stability investigation

Peptide drugs can be administered through injections (subcutaneous, intravenous etc.), via other routes such as trans buccal, nasal, inhaled, transdermal or orally [1-3,5,7,17]. When the peptide therapeutic is administered orally there are several enzymatic barriers that should be crossed to get to the potential target and thus it can be digested by numerous proteases. Consequently, proteolysis is one of the major elimination pathways for most peptide drugs.

Therefore, in early preclinical studies *in vitro* metabolic stability assays should be performed using incubations of the peptide of interest in individual proteases or complex matrices such as blood, plasma or serum. Moreover, cell cultures can be used to measure metabolic stability of the compound, i.e. in hepatocytes cells. During sample preparation experimental conditions can be optimized by evaluating and comparing different organic solvents to obtain an adequate extraction of the parent peptides and their metabolites and to optimize matrix effect. For example, in case of pepsin, pH should be reduced to activate the protease. Both kinetic information (*in vitro* intrinsic clearance and half-life) and degradation products can be determined when peptides are incubated with individual proteases or biological matrices in order to evaluate their stability. This information is used to guide structure modifications to improve peptide stability.

The typical matrices that are used for various species are:

- plasma/serum and blood to evaluate degradation in systemic circulation;
- GI fluids (simulated gastric fluid (SGF), simulated intestinal fluid (SIF)), intestine brush border membrane vesicles (BBMV), and intestine microsomes or S9 to examine GI stability and predict oral bioavailability;
- liver microsomes and hepatocytes to study liver metabolism by the various liver enzymes;
- kidney BBMV, microsomes, or homogenates to assess kidney degradation;
- tissue homogenates to examine tissue stability;

- assay media and formulation vehicles to ensure acceptable stability.

Recently, information was collected for about 500-600 human proteases in total and from these group about 300-400 are functional in the human body [45-48]. MEROPS [49,52], CutDB [50] and ENZYME databases [51] integrate available information about proteolytic sites and, consequently, about proteases, their cleavage sites, substrates and inhibitors. This information can be used to identify possible labile residues in the candidate peptide for individual proteases. But in drug discovery, the proteolytic enzymes for a specific peptide are not always known. Consequently, the sites of cleavage are typically identified using liquid chromatography–mass spectrometry (LC-MS) through elucidation of the metabolites structures.

1.2.2 Analytical techniques to perform metabolite identification of peptides

Drug metabolism properties are one major determinant that characterizes a successful drug candidate. Metabolic stability largely contributes to bioavailability of the active compound and thus the pharmacological response. Nowadays, many analytical techniques for the identification and structure elucidation of metabolites are available. This knowledge can help to understand aspects of bioavailability and it is crucially important to get this information as soon as possible during the drug discovery process.

The study of peptide drug metabolism is a complicated process which requires sophisticated analytical techniques. Selection of the most suitable technique generally requires a compromise among speed, selectivity and sensitivity. Currently, NMR, gas chromatography (GC), liquid chromatography including high performance liquid chromatography (HPLC), and capillary electrophoresis are widely used in metabolic profiling and targeted metabolite analysis for disease diagnosis and treatment, detection and identification of biomarkers and drug synthesis as well as metabolism investigations [53-55]. Moreover, analytical separation methods such as GC, HPLC coupled with mass spectrometry have been investigated for their application to metabolism research.

Mass spectrometry is an essential tool for efficient and reliable quantitative and qualitative analyses: metabolite formation and identification for small molecules, peptides and others. The characterization of metabolites in biological matrices becomes more challenging as the complexity of the matrix background increases. Mass spectrometry is particularly attractive because of the higher selectivity between similar peptides, higher accuracy and precision and lower requirement for specific reagents [36]. Mass spectrometry and tandem mass spectrometry (MSMS) experiments are the major tools used in protein and peptide identification. Mass spectrometers measure the mass/charge ratio of analytes. For protein studies, this can include intact proteins and protein complexes [37]. In case of top-down sequencing fragment ions produced by gas-phase activation of protein ions are measured [56-59]. In case of bottom-up sequencing peptides produced by enzymatic or chemical digestion of proteins (mass mapping) [39,60], and fragment ions produced by gas-phase activation of mass-selected peptide ions are measured [61]. The application of MS and MSMS to proteomics takes advantage of the vast and growing array of genome and protein data stored in databases. Protein identification by mass spectrometry requires an interplay between mass spectrometry instrumentation (how molecules are ionized, activated, and detected) and gas-phase peptide chemistry (which bonds are broken, at what rate, and how cleavage depends on factors such as peptide/protein charge state, size, composition, and sequence) [60].

Different types of mass spectrometers that serve different purposes are available to the DMPK (drug metabolism and pharmacokinetics) scientists. Originally triple stage quadrupole mass spectrometers had been developed to sensitively, specifically and precisely measure drugs and metabolites in biological matrices and enable qualitative analysis using scanning modes. These scanning modes are full scan, precursor ion scan, selected reaction monitoring, or neutral loss scan as survey experiments and product ion scan as dependent experiments [62]. For the MS techniques the accurate mass data obtained by using time-of-flight (TOF) mass spectrometers was a breakthrough in metabolite profiling. Therefore, typically triple stage quadrupole is used for the quantitative analysis and quadropole time-of-flight (QTOF) and ion trap (IT)/Orbitrap mass spectrometers are used for qualitative analysis. High resolution (resolving power RP >30,000) with good mass accuracy (<5 ppm) can be reached by two mass spectrometer type of instruments: TOF-MS and Fourier transform (FT)- MS, including the FT ion cyclotron resonance and the Orbitrap [70]. One of the main difference between the two types of instrument used for the peptide incubation data is the duty cycle [63].

More recently, liquid chromatography coupled with high-resolution mass spectrometry (LC-HRMS) has gained an important role [38] for both quantitative [64] and qualitative analysis in DMPK studies [62] specially, when metabolites should be separated and identified with a high degree of certainty in complex biological matrices [37, 39,60,61]. One of the main advantages of HRMS is that it almost does not need to be tuned and large numbers of analytes can be analyzed simultaneously. Liquid chromatography is extremely important in analyte preconcentration and sample cleanup. Significant progress has been achieved in separation efficiency with the development of core-shell technology, resulting in LC peak widths of only a few seconds [65]. Additionally, improvements in chromatographic separation such as ultra-high-performance liquid chromatography (UPLC) have generally led to higher chromatographic resolution via sharper peaks and correspondingly higher MS sensitivity [66, 70].

Typically, LC-HRMS is applied using data-dependent acquisition (DDA). In DDA, precursor ions are selected based on their abundances and are serially fragmented. Although commonly used, DDA has inherent limitations, such as stochastic and irreproducible precursor ion selection, undersampling and long instrument cycle times. The higher resolution, improved mass accuracy and sequencing speed of the most recent mass spectrometers partially reduce these problems. Also, the introduction of a preferred list of ions that represents the potential products formed can help in having a more reliable precursor ion selection. Unbiased data-independent acquisition (DIA) strategies have also been developed to overcome the limitations of DDA and improve reproducibility. DIA approaches feature parallel fragmentation of multiple precursor ions, regardless of intensity or other characteristics, resulting in complex but comprehensive production of data. DIA methods enable the acquisition of a complete, unbiased sample record, enhancing quantification reproducibility. During MS^E (where E is collision energy) acquisition (one of the most common DIA methods), alternating MS scans are collected at low and high collision energy, providing information on precursor and fragment ions, respectively. Chromatographic coelution profiles of precursor ions and their corresponding fragment ions can be used to generate deconvoluted product-ion spectra for each precursor, which can be subsequently searched against databases. Thus, in contrast to DDA-based methods, which are intrinsically limited by scan time, DIA methods are theoretically limited only by system peak capacity [57].

An established DIA approach to data collection, such as ‘all-in-one’ fragmentation or MS^E [66], employs a rapid alternation between two full scan MS functions. The first scan function applies a low collision energy which results in precursor ion spectra (drug and metabolites), and the second scan function acquires data at high collision energy resulting in almost simultaneous acquisition of high resolution fragment ion (FI) spectra. The use of QTOF platforms with UPLC provides well-resolved peaks and in most cases the predominant FIs can be associated with a single matching precursor ion [66]. An extension of the MS^E approach was enabled by the introduction of ion mobility (IM) functionality into mass spectrometers. Briefly, IM-MS is a two-dimensional separation technique that separates ions in a dimension related to structure (charge-to-surface area ratio) as a function of their collision cross section (CCS) and subsequently in a second dimension according to the mass-to-charge ratio [56]. CCS represents the area of the ion available for collisions with neutral molecules in the gas phase. Many forms of ion mobility exist: high field asymmetric waveform ion mobility (FAIMS), differential ion mobility, traveling wave ion mobility (TWIMS) and uniform field ion mobility (IMS). A specific feature of ion mobility spectroscopy, when coupled with mass spectrometry and a post-IMS collision cell, is that FIs can be correlated with their precursor ions based on a shared drift time to generate IM-resolved spectra. This approach is particularly useful in complex samples [60] and $HDMS^E$ approaches have been applied to a variety of biological, pharmaceutical and environmental scenarios [53]. IMS provides an additional dimension of separation by improving system peak capacity while concomitantly reducing chimeric and composite interferences [56,57]. A key aspect of the combination of an IMS separation (typically occurring in the millisecond time-frame) and MS detection (typically occurring in the microsecond time frame) is that it allows an additional separation step to be obtained on the MS time-scale (e.g., in addition to liquid chromatography), without compromising the speed of MS detection [67].

1.2.3 Semi-automated tools for metabolite identification of peptides based on MS data

Technological advances in mass spectrometry such as accurate mass high resolution instrumentation have fundamentally changed the approach to systematic metabolite identification over the past decade [47]. Notwithstanding the sensitivity of these platforms, and the quality of the data which can be generated, their usage in drug metabolite identification can be a time-consuming task, several semi-automated tools were developed for MS peptide data interpretation. These approaches include four main groups:

1. database searching (SEQUEST, MASCOT, etc.);

2. de novo peptide sequencing (PEAKS, PepNovo, etc.);

3. peptide sequence tagging (GutenTag);

4. consensus of multiple search engines (Scaffold) [68].

These MS-based proteomics approaches have difficulties with sequencing cyclic peptides without prior linearization and they are limited to the 20 standard amino acids [2, 46,47].

Therefore, other semi-automated cheminformatic approaches were developed to assist the structural assignment of metabolites based on known parent structures such as:

- BiopharmaLynx [69, 71];

- MetabolitePilot [72];

- Mass-MetaSite (MMS) [72-78].

These software tools are able to propose metabolite structures based on the combination of metabolite prediction and interrogation of analytical mass spectrometric data.

BiopharmaLynx is a new informatics package for vendor-specific software within the MassLynx approach [69, 71] that can analyze LC-MS data during peptide mapping studies. The basis of the MassLynx software algorithm is designed to compare each analyte sample with a control sample and perform correlation between retention time, *m/z* value, intensity and components from alternative detection technologies (e.g. diode array UV). This comparison allows to filter matrix-related peaks, therefore, excludes the false positive ones. BiopharmaLynx automatically annotates detected peptides structures related to the identified LC-MS peaks using accurate mass assignments and utilizes algorithms for filtering compound-related ions from endogenous matrix ions to identify only ions that are related to parent compound. It can recognize modified peptides in the analyzed sample [69, 71].

MetabolitePilot is a vendor-specific software package used in both pre- and post-data acquisition processes. Before acquisition during parent structure import this software generates inclusion list based on theoretical fragmentation of parent molecule. It includes predicted biotransformation reactions and any potential cleavage metabolites. This list is used to preferentially trigger MSMS data collection during acquisition for all included drug-related sample components, but also can trigger on any unexpected peaks above a preset intensity threshold. In post-acquisition phase MetabolitePilot is used to process the data in batches and extract metabolite information through comparison of up to five control(s) and a treated sample. The following information and features are used to identify potential metabolites: common product ions, neutral losses, isotope pattern and/or multiple mass defect filters. Also, it performs additional filtering of the false positives based on mass accuracy, score, retention time and MSMS similarity and selection of the major metabolites of interest based on analog and MS peak areas. Proposed structures should be checked by expert. Finally, each of the elucidated structures is scored to measure the confidence in a given interpretation [72].

During this study we used Mass-MetaSite [72-78] a vendor neutral approach that uses LC-MS data from peptide metabolic stability experiments to determine the specific metabolic sites for processing DIA (MS^E and $HDMS^E$) and DDA data. This tool can process datasets from both small molecule and peptide metabolic stability experiments to determine the specific sites of metabolism for small molecules and metabolic cleavage sites for peptides and then store the results in a chemically aware database (WebMetabase, WMB), where chemical structure-based searches can be performed by structure and/or substructures. Mass-MetaSite uses as inputs the 2D structure of the compound together with control and treated sample data files. The data can be processed sample-by-sample manually or in a batch mode with an automatic processing of a set of sample files. MMS data processing is shown in Figure 5.

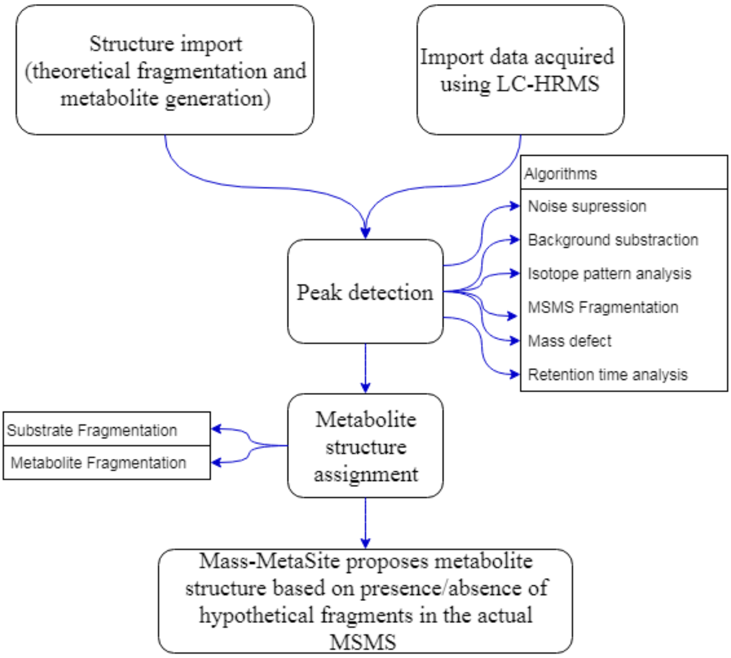

Structure import (theoretical fragmentation and metabolite generation)
Import data acquired using LC-HRMS
Algorithms
Noise supression
Background substraction
Isotope pattern analysis
MSMS Fragmentation
Mass defect
Retention time analysis
Peak detection
Substrate Fragmentation
Metabolite Fragmentation
Metabolite structure assignment
Mass-MetaSite proposes metabolite structure based on presence/absence of hypothetical fragments in the actual MSMS

Figure 5. Mass-MetaSite metabolites elucidation process

The data processing involves two steps (Figure 5). Step-1 consists of automatic detection of the chromatographic peaks related to the parent compound, i.e. metabolites. The methodology for peptides does not differ from the one described for small molecules [72]. Step-2 consists of structure elucidation of the potential metabolites based on the fragmentation pattern for each peak detected. Once the list of potential chromatographic peaks has been selected (step-1), Mass-MetaSite compares the *m/z* associated with each peak to all the possible theoretical metabolites based on a list of included biotransformation reactions [72]. In this study, the only transformation of interest selected was the hydrolysis of amide bonds.

The overall principle for the structural elucidation of metabolites is a comparison of fragment ions obtained from the parent (assigned from the incubation time t = 0 sample) and the ones from the metabolites (t = incubation time) and then identifying mass shifts corresponding to the mass of the metabolite or common neutral losses [76].

In addition to the above comparative fragmentation analysis, the fragmentation of the metabolite without comparison to the parent molecule is considered. This fragmentation strategy is most advantageous in the case of cyclic peptides where the metabolite could be a linear peptide (the amide hydrolysis is occurring in the cycle) and fragmentation can be significantly different compared to the parent one. Fragmenting all the metabolite structures to the same extent as the substrate takes a prohibitive amount of computational time; therefore, the number of bonds that can be broken to generate metabolite fragments has been limited to 1.

A score is assigned to each peptide metabolite based on the peak intensity and the number of matches/mismatches between the theoretical fragment *m/z* value and the *m/z* value observed in the MSMS spectrum [78]. Once Mass-MetaSite results have been uploaded into a database system like WebMetabase, they should be manually checked and approved by the expert.

Mass-MetaSite can process the experimental data of peptide substrate structures up to 4000 Da in molecular. This limitation is related to the fact that maximum monoisotopic mass is used for the peak detection.

During this study we used Mass-MetaSite to HRMS DDA and DIA data from *in vitro* incubation samples, to determine specific metabolic cleavage sites of peptides which involves reading peptide MSMS data, finding the chromatographic peaks related to the parent compound and elucidating the structure of the metabolites. Also, we used MMS to process MS^E and HDMS^E data to compare the accuracy and quality of metabolites elucidation for small molecules and peptides.

1.3 *In silico* cleavage site prediction tools

1.3.1 Proteases catalytic mechanisms

Recently, several protease classification systems were published, and they are shown in Figure 6.

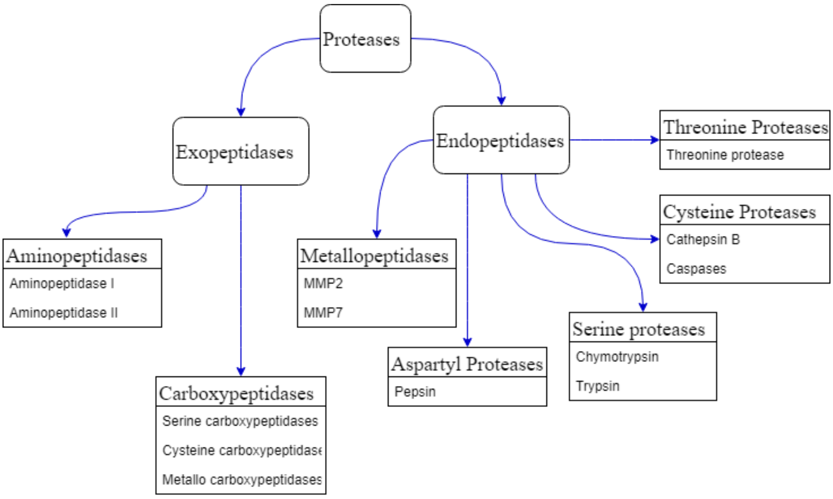


Figure 6. Proteases classification

Proteases were classified in two groups: endoproteases and exoproteases, depending on the location of the target in the sequence. Endopeptidase action is directed on internal peptide bond, exopeptidase action is directed by the N-termini or C-termini (amino- and carboxypeptidases, respectively).

Proteases fall into six main classes with respect to their catalytic mechanism:

1. serine proteases;
2. cysteine proteases;
3. threonine proteases;

4. aspartic proteases;
5. glutamic proteases (not found in the mammals);
6. the metalloproteases.

For the remaining proteases the catalytic mechanism is known [49,79]. Proteases can be further classified into families that contain related sequences, which are clustered into clans that contain related tertiary structures. Metalloproteases and serine proteases are the most densely populated classes, respectively, followed by cysteine proteases, whereas threonine and aspartic proteases contain lower number of members, respectively [79].

Proteases in each class cleave peptide bonds through different catalytic mechanisms. Aspartic, glutamic, and metalloproteases utilize an activated water molecule as a nucleophile to attack the peptide bond of the substrate. Cysteine, serine, and threonine proteases, the nucleophile is an amino acid residue (Cys, Ser, or Thr, respectively) located in the active site that participates in covalent catalysis [79].

The primary determinant of protease specificity is the architecture of the protease active site because of its ability to interact with the amino acid residues surrounding scissile bond in the substrate. Proteolytic cleavage of peptides is directed by short amino acid motifs, from two to eight amino acids around the scissile bond that binds to the specific pocket that defines proteases specificity. The amino acids surrounding the cleavage site are called P4-P3-P2-P1-P1'-P2'-P3'-P4' with cleavage site located between P1-P1' [81,82]. The selectivity of the protease can vary depending on the amount of the preferred specific residues around the cleavage site of the substrate from generic proteases who discriminate only one residue in the cleavage site to highly specific when eight amino acids are important. Although the specificity could also depend on exosites and allosteric sites [80]. Active sites for cysteine, serine, aspartic proteases and metallopeptidases are presented in Figure 7.

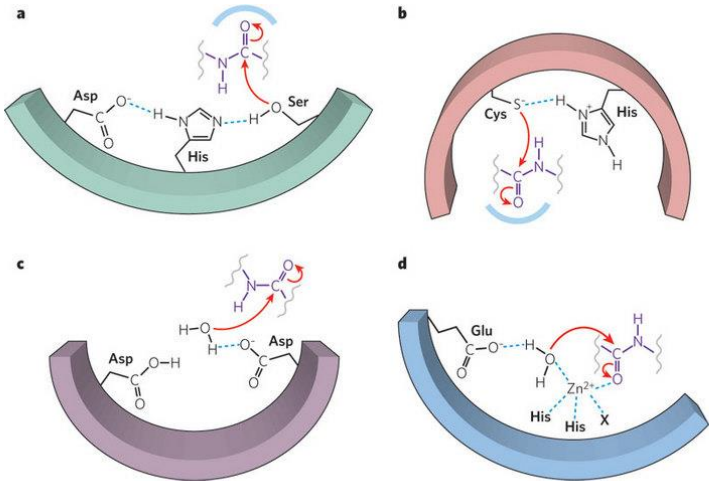


Figure 7: Protease mechanisms for serine proteases (a); cysteine proteases (b); aspartyl proteases (c); and metalloproteases (d). © 2009 Nature Publishing Group Erez, E., Fass, D., & Bibi, E. How intramembrane proteases bury hydrolytic reactions in the membrane. Nature 459, 371–378 (2009).

Specificity mechanism depending on serine protease for chymotrypsin, trypsin and elastase are illustrated on Figure 8.

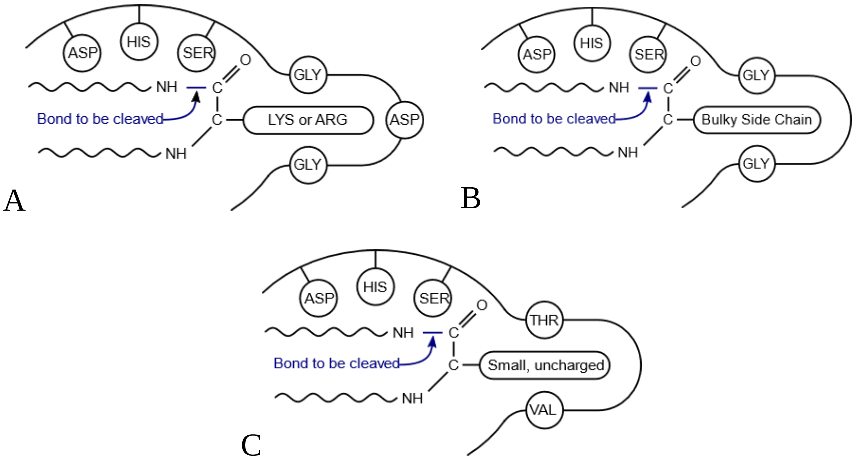


Figure 8. Specificity mechanism for serine proteases: A) Chymotrypsin; B) Trypsin; C) Elastase

Most proteolytic enzymes cleave α-peptide bonds between naturally occurring amino acids, but there are some proteases that perform slightly different reactions. Thus, a large group of enzymes known as DUBs (deubiquitylating enzymes) can hydrolyze isopeptide bonds in ubiquitin and ubiquitin-like protein conjugates; γ-glutamyl hydrolase and glutamate carboxypeptidase target γ-glutamyl bonds; γ-glutamyltransferases both transfer and cleave peptide bonds; and intramolecular autoproteases (such as nucleoporin and polycystin-1) hydrolyze only a single bond on their own polypeptide chain but then lose their proteolytic activity [79].

1.3.2 Cleavage data databases

Nowadays, the information about proteases and their cleavage sites is widely spread across publications and databases but it is not always suitable for fast automatic computed searches. Following databases integrate available information about proteolytic sites and, consequently, about proteases, their cleavage sites, substrates and inhibitors:

Table 3. Databases that integrate available information about proteases, their cleavage sites, substrates and inhibitors.

ENZYME database	Bairoch A. The ENZYME database in 2000. Nucleic Acids Res. 2000;28(1):304-305. doi:10.1093/nar/28.1.304.
CutDB	Igarashi Y, Eroshkin A, Gramatikova S, et al. CutDB: A proteolytic event database. *Nucleic Acids Res*. 2007;35(SUPPL. 1):546-549. doi:10.1093/nar/gkl813

Table 3. Databases that integrate available information about proteases, their cleavage sites, substrates and inhibitors.

MEROPS	Rawlings ND, Waller M, Barrett AJ, Bateman A. MEROPS: the database of proteolytic enzymes , their substrates and inhibitors. 2014;42(October 2013):503-509. doi:10.1093/nar/ gkt953. Rawlings ND, Barrett AJ, Thomas PD, Huang X, Bateman A, Finn RD. The MEROPS database of proteolytic enzymes, their substrates and inhibitors in 2017 and a comparison with peptidases in the PANTHER database. Nucleic Acids Res. 2018;46(D1): D624-D632. doi:10.1093/nar/gkx1134.
Proteasix	Klein J, Eales J, Zürbig P, Vlahou A, Mischak H, Stevens R. Proteasix: A tool for automated and large-scale prediction of proteases involved in naturally occurring peptide generation. *Proteomics*. 2013;13(7):1077-1082. doi:10.1002/pmic. 201200493.
PMAP	Igarashi Y, Eroshkin A, Gramatikova S, et al. CutDB: A proteolytic event database. *Nucleic Acids Res*. 2007;35(SUPPL. 1):546-549. doi:10.1093/nar/gkl813 Igarashi Y, Heureux E, Doctor KS, et al. PMAP: databases for analyzing proteolytic events and pathways. *Nucleic Acids Res*. 2009;37(suppl_1): D611-D618. doi:10.1093/nar/gkn683.

This information can be used to identify possible labile residues in the candidate peptide drug for individual proteases. Most of the information in these databases is limited to 20 standard amino acids [46].

ENZYME nomenclature database is a repository that contains the nomenclature of enzymes referenced by Enzyme Commission (EC) number and based on the recommendations of the Nomenclature Committee of the International Union of Biochemistry and Molecular Biology (IUBMB) [51].

The CutDB database [50] was one of the first publicly available databases created to collect information regarding individual cleavage sites reported in the literature. The database contains information about more than 5300 annotated proteolytic events that occur within 1702 protein substrates. These cleavage sites were registered for 164 metallopeptidases, 180 serine, 108 cysteine and 61 aspartic proteases.

The MEROPS is a publicly available peptidase database [49, 52] that integrates the available information about proteases from different organisms, their experimentally identified or predicted SoCs with their sequences, substrates and inhibitors. At the moment of writing, it contains over 900,000 sequences of the peptidases and more than 90,000 substrate cleavages and over 80% of them were mapped to UniProt database, the remaining ones were represented by cleavage sites in synthetic substrates. Peptidases are hierarchically classified in MEROPS. Each peptidase is annotated by primary substrate binding sites (though not necessarily secondary binding sites, known also as 'exosites') and the catalytic residues. Peptidase normally corresponds to a structural domain, and some proteins contain more than one peptidase domain. Each substrate is described by: a) the name, b) the UniProt accession number, c) the peptidase known to cleave that substrate with a link to the summary for that peptidase, and a count of cleavages performed by each peptidase, d) the residue number of the amino acid in the P1 position, e) the name of the peptidase responsible, f) the residue range of the substrate used in the experiment compared to the complete coding sequence that is presented in the UniProt entry, whether the cleavage is thought to be physiological or not; g) how the cleavage was determined (e.g. MS for mass spectroscopy etc.), h) a comment describing the purpose of the cleavage and a reference. Different types of searches can be performed in the database: sequence similarity search for peptidases, substrates and inhibitors with BLASTP, PSI-BLAST, FASTA, HMMER; search of the peptidase by name, MEROPS identifier, other identifiers such as UniprotKB, gene name, by family or clan name; search by specificity can be performed using cleavage site sequence, substrate name, substrate UniProt accession, inhibitor, etc. User can add sequence of characterized proteins and hypothetical sequences from organisms that are of evolutionary, medical or economic interest.

Proteasix is publicly available database of cleavage sites that contains information about 3500 protease/cleavage sites combinations from 191 different human proteases. In total over 1700 unique protease/substrate pairs were described in the database. It was built using information collected in CutDB, Uniprot database and available literature. Each protease was annotated with cleavage site P4-P4' and Swiss-PROT identifier number. Each cleavage site was aligned with Swiss-PROT substrate sequence. The user can perform a search of the cleavage sites using Swiss-PROT identifier or by name of substrate peptide, peptide substrate's start and end. The output of the search is the list of the proteases associated with predicted cleavage sites. Mentioned search is performed through alignment of the input peptide substrate sequence to Swiss-PROT sequences.

PMAP (proteolysis MAP) is a website that helps to improve understanding regarding proteolytic networks and pathways and proteolysis reasons. PMAP is connected to five databases (CutDB ProteaseDB, SubstrateDB, ProfileDB and PathwayDB) and thus links information from these databases together, where CutDB has information on more than 5000 proteolytic events. ProteaseDB contains basic information for a set of 150 human proteases and more than 45000 proteases acquired from MEROPS. The information is stored in a MySQL database and presented as a Molecule Page on a web server that displays a comprehensive set of annotations on 15 different features of proteases such as: a) predictions of PFAM domain structure [85], b) secondary structure, transmembrane regions, c) signal peptides and disordered regions and each page is linked to external sites (e.g. MEROPS, PDB, PubMed, etc.) from which the data was collected. SubstrateDB contains molecular information on documented protease substrates in CutDB. Data was extracted from MEROPS, human protein reference database, UniProt and original articles. For each cleavage site the primary sequence of the protein substrate with highlighted residues around cleavage site is described. Moreover, the following information is stored for each proteolytic event: a) molecular identity of the protease linked to ProteaseDB, b) substrate linked to

SubstrateDB, c) reference in the Literature Track, d) and other features associated if available (e.g. the method by which the event was detected, the potential consequences of the event, relevant cofactors, associated pathways, cell compartments where the event takes place, cell lines where the event is observed and information linking the event to any process or disease). Any registered user can curate the content of CutDB by adding new events, fixing errors or adding comments [49,84]. ProfileDB was generated to collect information of the substrate recognition specificity of proteases. PathwayDB uses data from four databases described above and comprise information about known pathways and hypothetical pathways reports in the literature [84].

Information stored in these databases can be used to identify possible labile residues in the candidate peptide for individual proteases. Although useful, these databases still have several limitations. None of the resources allows to perform an automatic search in batches, thus each Site of Cleavage (SoC) should be queried manually, except of Proteasix. But Proteasix contains information on a reduced set of proteases. Furthermore, none of the resources allows to add new experimental information in an automatic way to enrich the database information. Finally, these databases are publicly available and thus cannot be enriched with private data.

1.3.3 WebMetabase Database

In this study we planned to develop a new approach that uses LC-HRMS DDA, DIA (MS^E or $HDMS^E$) data from peptide metabolic stability experiments or data coming from external sources such as MEROPS database to determine the specific metabolic cleavage sites. The peptide mode of Mass-MetaSite will be used to process the MS data and elucidate the metabolites structures and upload results into WebMetabase. Metabolites structures extracted from MEROPS will be directly uploaded into WMB. Therefore, this system should be able to combine data from multiple data sources. When data is stored in a chemically aware database (e.g. WebMetabase), each substrate and cleavage site is annotated by residues or structural blocks (SBs) contained in the sequence and described as a combination of pharmacophoric and/or physicochemical properties of each SB contained in the sequence.

For this purpose, we planned to use molecular descriptors (Volsurf descriptors [109,110] and SHOP descriptors [104-108], respectively). Therefore, the proposed methodology should be able to handle synthetic amino acid. Also, each stored residue is annotated with similarity score calculated based on pharmacophoric and/or physicochemical properties.

Moreover, this new way to annotate the information should also include a system to perform an exact substructure and similarity-based substructure search without being limited to any type of peptides and/or amino acid (cyclic/linear, natural/synthetic). Therefore, the methodology could be used to perform a structural search on the exact sequences to identify the experiments where a certain bond of interest participated in a metabolic reaction. In addition, the system should also be able to perform searches based on similarity of the molecular descriptors identifying the bond participated in a metabolic reaction that is similar to the bond of interest. The output of the searches should be a list of experiments that fulfill the search criteria. Additionally, the proposed methodology should be able to handle unnatural amino acid and/or cyclic peptides. The system should be linked to the software assisted metabolite structure elucidation based on MS data (e.g. Mass-MetaSite), so that the database can be automatically enriched with the new experiments.

1.3.4 Available *in silico* prediction tools

Predicting possible sites of cleavage for individual proteases is an important task to be completed during drug-design process of peptide therapeutics. Such information could be used to improve their stability and bioavailability as a promising drug. Currently, mass spectrometry techniques or peptide libraries profiling are typically used to identify possible sites of cleavage in peptide drugs. However, in most cases experimental identification of protease cleavage sites is a difficult, labor-intensive and time-consuming task and requires access to specialised equipment. Moreover, new peptide molecule synthesis is labor-intensive and time-consuming. In contrast to experimental methods, *in silico* prediction of proteolytic sites has been recognised as a useful alternative approach to provide valuable knowledge on probable protease-peptide drug interaction relationships. Efficient computational tools would reduce the number of experiments and would help to improve peptide drug structure before synthesis.

Recently, several bioinformatics tools were implemented to identify possible proteolytic sites and proteases that can probably cleave the peptide of interest. These approaches use as input data extracted from databases such as MEROPS [49,52], CutDB [50], collect information from the available literature and/or experimental data. These approaches can be classified in two main groups:

1. tools that provide general prediction for proteases from different classes (SitePrediction [86], PROSPER [101], PROSPERous [87], PoPS [88], ExPaSy [89, 90]);

2. tools that provide prediction only for the specific proteases or the specific classes of proteases (CasCleave 2.0 [91], GraBCas [92], CASVM [93], Pripper [94], CasPredictor [95]).

These approaches use protease specificity prediction models and depending on the way how models were developed they can be classified in four main groups:

1. Sequence-based approaches (ExPaSy);

2. Approaches that perform prediction using position-specific scoring matrix (PSSM) for individual proteases (GraBCas, CasPredictor, PoPS, SitePrediction);

3. Approaches that use predictive models trained on sets of cleavage site specific descriptors (PROSPER, PROSPERous, Pripper, CasCleave 2.0, CASVM);

4. Approaches that combine methods explained (Proteasix [83]).

The first group contains sequence-based approaches that use exact sequence motifs matching to a known cleavage site. The main limitation of these approaches is that they are restricted to the set of the known sites.

Approaches in the second group perform prediction of possible cleavage sites using PSSM for individual proteases based on the available information regarding cleavage sites of different proteases. The main limitation of these type of tools is that they are restricted by amino acids used to build scoring matrix and thus most of the cases limited to 20 standard amino acids and cannot perform prediction on synthetic peptides with unnatural amino acids.

In the third group approaches each proteolytic site is represented as a set of cleavage site specific descriptors that defines the identity of each residue in the proteolytic site sequence. Different types of descriptors can be used: descriptors that explain physical, physicochemical, pharmacophoric properties, secondary structure of the cleavage site, solubility of the sequence around cleavage site and others. At the next step these approaches use different machine learning algorithm such as support vector machine learning algorithm (CASVM, Pripper, CasCleave 2.0, PROSPER, PROSPERous), logistic regression (PROSPERous), neural networks and others to train prediction models based on these descriptors. These models can use different size of local window around cleavage site, typically from two to sixteen residues.

Finally, approaches in the fourth group combine methods explained above. For example, Proteasix can perform matching against the collection of the known cleavage sites from the literature and calculate probability of cleavage event appearance based on MEROPS specificity matrix.

Since these tools use input data from databases such as MEROPS to develop models and identify possible labile residues in the candidate peptide drug for individual proteases, they still have several limitations. For example, none of the available resources allow to add new information in an automatic way to enrich the database content and tune the models. These tools do not provide a methodology for user to train their own models using private experimental data. Finally, these tools contain models trained for specific proteases and are not able to perform model developments for complex matrices (e.g. plasma, serum, blood etc.) or develop models for a specific group of the peptides. The later group of models can be particularly usefull for the development of the synthetic analogues for a peptide of interest (e.g. analogues of LHRH, analogues of somatostatin etc.).

1.3.5 WebMetabase Analysis Tools

The availability of the database described above enables the third aim of the present work that is to be able to perform frequency analysis to discover the most frequent metabolically labile amide bonds. These frequency analysis results can be used to understand cleavage site appearance rules for the specific peptide family or for specific experimental condition (i.e. individual protease) within this database, like in the methodologies from the first group of models described above. Moreover, these results can be used to train cleavage site prediction models for individual proteases or complex matrices like in the third group of approaches. For this purpose, we planned to use and compare several machine learning algorithms since it was demonstrated in the literature that these methods are highly effective when applied in *in silico* tools for prediction of the sites of cleavage [87, 91, 93, 101].

Since we plan to annotate each amino acid and each cleavage site by its physicochemical and pharmacophoric properties using molecular descriptors, we can use these descriptors for the model training. In this case there would be no limitations related to the amino acids used for the training as it happens when the exact sequences of amino acids are used for the training. Also, our goal is to develop a system that can be linked to the software assisted metabolite structure elucidation based on MS data, consequently the database can be automatically enriched with the new experimental data and derived rules can be refined to tune the system for the experimental conditions and/or peptide families of interest.

2. RESULTS AND DISCUSSIONS

Peptides therapeutics are gaining a significant role on the pharmaceutical market due to high selectivity and efficacy. However, one of the main aspects that usually have to be optimized in the case of peptide-based drugs is the low bioavailability and instability due to protease activity. Therefore, it is crucially important to identify primary metabolites of the peptides for incubations with individual proteases or complex matrices such as plasma in an efficient manner. In addition, *in silico* tools able to predict possible cleavage sites for various incubation conditions are also needed. This knowledge can be applied during drug development process to better understand when and where structural modifications are required to improve peptide ADME properties: stability and bioavailability. Currently, databases such as MEROPS or Proteasix integrate information regarding many proteases, their substrates, cleavage sites and inhibitors, collected from available literature and other public sources. This data is used in several publicly available *in silico* software tools such as PeptideCutter, PROSPER or PROSPERous to predict the Site of Cleavage (SoC). These approaches can help to design a peptide drug with increased stability against proteolysis. Nevertheless, the databases and tools described above have several limitations:

1. Databases:
    a. Databases used for the model training dataset preparation cannot be enriched with new experimental data in an automatic way.
    b. Information for some type of peptides is not well captured i.e. peptides with nonstandard monomers, peptides with single or multiple cycling bonds.
2. Model tools:
    a. A limited number of proteases have been modelled in the available tools.
    b. There are no prediction models trained for a complex matrix such as plasma.
    c. Models cannot be refined and tuned with new experimental information.

d. These tools are publicly available and therefore there is no possibility to build private models for internal use.
e. Several tools cannot perform prediction on synthetic peptides that contain unnatural amino acids or peptides with cyclic structure.

Consequently, our main aim for this thesis is to introduce and validate new workflow to predict and rank potential cleavage sites for a new candidate peptide-based drug without any limitations on peptide structure and no limitations in data source using the derived specificity rules for specific protease or complex matrix. In Figure 9 the main workflow followed in this thesis is represented.

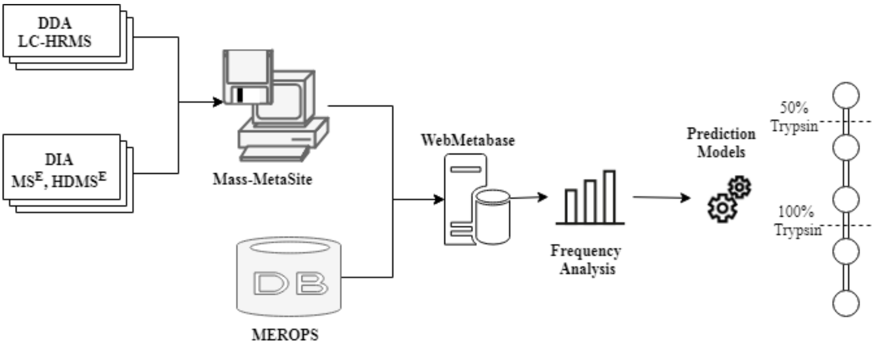


Figure 9. Developed workflow to perform cleavage site prediction for peptide-based drug candidates

This workflow contains several steps:

1. In the first step we implemented a methodology based on Mass-MetaSite (MMS) to process coupled to liquid chromatography - high-resolution mass spectrometry data (LC-HRMS) from *in vitro* incubation samples (Figure 9) (***Articles 1, 2, 3***).

2. In the second step the system is used to store the results in the chemically aware database WMB (Figure 9) (***Articles 1, 2, 3, 4***).

3. In the third step, we implemented an algorithm that performs frequency analysis (FA) based on exact match of residue structures in the site of cleavage (so-called simple frequency analysis, SFA) (Figure 9) (***Articles 1, 2, 3***).

4. In the fourth step, we implemented an algorithm that performs similarity frequency analysis (SimFA). This analysis reveals the group of metabolically labile amide bonds defined by similar pharmacophoric or physicochemical properties towards the specific proteases/specific media (Figure 9) (***Article 4***).

5. In the fifth step, we implemented system that used created database to train cleavage site prediction models using different machine learning algorithms based on molecular properties of the identified cleavage sites (Figure 9) (***Article 5***).

2.1 First step

To complete these goals at the first step we implemented a methodology based on Mass-MetaSite (MMS) to process data-dependent acquisition (DDA), data-independent acquisition (DIA) MS^E and ion mobility-MS (high definition MS^E, $HDMS^E$) data from *in vitro* incubation samples (Figure 9). This developed methodology enables the elucidation of metabolite structures and thus allows to identify the cleavage sites. The step of the data processing involves reading peptide MSMS data, finding the chromatographic peaks related to the parent compound and elucidating the structure of the metabolites based on the fragment ions for parent and metabolite structures.

In the first article we used the implemented approach to process DDA LC-HRMS analytical data collected for a set of fourteen peptide drugs (linear/cyclic, containing natural/unnatural amino acids) and four substrate peptides incubated with different proteolytic media: trypsin, chymotrypsin, pepsin, pancreatic elastase, dipeptidyl peptidase-4 (DPP4) and neprilysin (NEP). This peptide set included two families of analogues: the LHRH peptide analogues that includes LHRH peptide and five synthetic analogues and glucagon-like peptide-1 (GLP-1) analogues group including GLP-1 and three synthetic analogues. The first group was used to investigate the effect of small chemical/monomer changes in the peptide structure with respect to the matrix catalyzed activity for chymotrypsin, trypsin, elastase and pepsin. In the second group the effect of small chemical changes in GLP-1 analogues group with respect to the proteases DPP-4 and NEP was also investigated.

During the metabolite identification study in total 132 metabolites were found from the various *in vitro* conditions tested.

In the second article a metabolite identification study was performed using a peptide set that included eight compounds: one commercially available - somatostatin - and seven synthetic analogues. All test compounds were incubated in serum. All eight peptides were cyclic and seven of them had unnatural amino acids. These peptides were used to investigate the effect of small chemical/monomer changes in the peptide structure with respect to the serum catalyzed activity. During metabolite identification 17 metabolites were annotated in the database.

In the third article DIA MS^E and $HDMS^E$ data for GLP-1 and verapamil incubated with rat hepatocytes were processed using the implemented workflow to compare accuracy and quality of data received using these DIA methods. A total of 7 metabolites were found for GLP-1. There was an agreement between the metabolites identified in these experiments and those reported in the literature with most of the rat metabolites being found by both unsupervised DIA methods [96].

All these multiple MS data sources were analyzed with MMS to find metabolites and determine their structures. All the metabolites identified were produced by amide hydrolysis and were checked manually and considered as reliable because the fragmentation was adequate, isotope pattern was as expected, the *m/z* small differences between the *m/z* of observed and theoretical, and the mass score was high.

In this way we have developed a new methodology that is able to process data from many different MS based analysis systems (DIA, DDA) for compound sets containing cyclic/linear and natural/non-natural peptides.

2.2 Second step

In the second step the system was used to store the results in a chemically aware database WebMetabase (Figure 9). The results uploaded into WMB are always followed by consolidation of all these data (cluster metabolites from different experimental conditions, i.e. incubation times from the same experiment). Furthermore, we added a new method that processes the information from external sources such as MEROPS database. The developed approach annotates each peptide by the structural blocks (SBs) defined as the structural fragments between amide bonds and their connectivity. The annotation of the peptide information in this manner enables doing a chemically aware exact substructure search. Moreover, each SB is annotated by its physicochemical and pharmacophoric properties that are computed by the validated Volsurf [109,110] and SHOP [104-108] molecular descriptors, respectively. Also, they are characterized by an index that describes the similarity between a SB and each one of the other SBs in the database. Thus, each peptide and cleavage site sequence are represented in the database as a combination of Volsurf and/or SHOP descriptors of the residues contained in the sequence. This type of annotation enables performing a similarity-based substructure search inside the database. The designed annotated system allows, that the searches are not limited to any type of peptides and/or amino acid (cyclic/linear, natural/unnatural) and do not consider the theoretical mass spectrum or even sequence alignment. The annotation system in the database described above is represented in Figure 10.

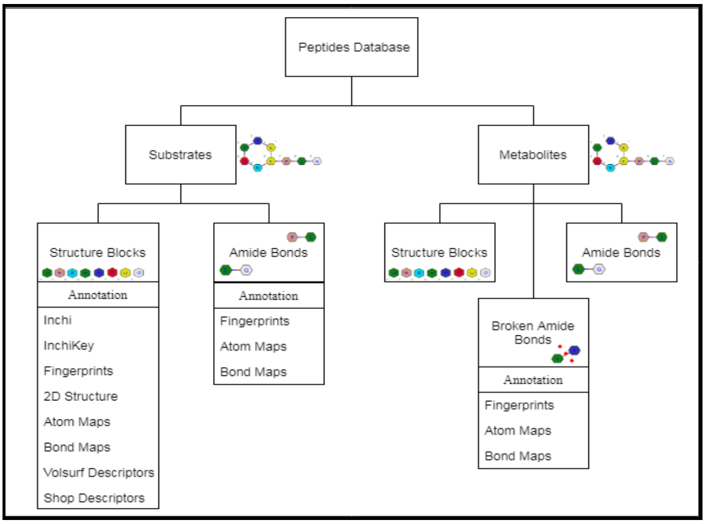


Figure 10. WebMetabase annotation system.

In the first article information regarding all identified metabolites was annotated into database. We compared the percent of remaining parent peptide with respect to the time for all investigated peptides to compute the clearance for each case. The smallest peptide, oxytocin, (molecular weight (MW) 1007.187) was digested slower than the biggest peptide, calcitonin (MW 3429.713) by the chymotrypsin protease. In the trypsin incubation oxytocin was hydrolyzed slower than all other peptides. These results agreed with the available literature published regarding incubations in gastrointestinal fluid (GIF) and simulated intestinal fluid (SIF) [85].

In addition, we examined the effect of small chemical changes in peptide structures for a) LHRH analogues group along with natural gonadorelin and b) GLP-1 analogues group along with GLP-1 peptide. All LHRH analogues were digested at similar rates by both chymotrypsin and trypsin except for leuprolide, which was hydrolyzed slower in all proteolytic media. This may be due to the replacement of Gly6 in gonadorelin with the non-natural amino acid, D-Leu to form leuprolide. In GLP-1 analogues group we found that DPP-4 and NEP hydrolyzed exenatide slower as compared to liraglutide and taspoglutide. This may be related to the substitution of Lys34 for arginine in exenatide and the addition of a C16 fatty acid at the ε-amino group of Lys26 using a γ-glutamic acid spacer in liraglutide. Liraglutide was digested faster than the other compounds in both matrices. For liraglutide, NEP acted slower as compared to DPP-4.

In the second article information regarding all identified metabolites was annotated into database. All compounds from the dataset were hydrolyzed with the different velocity. Five compounds were more stable than somatostatin, this may be due to the replacement of both Phe7 to Msa7 and Trp8 to D-Trp8. One of the compounds was digested significantly slower than somatostatin due to the replacement of both Cys3 and Cys14 to D-Cys.

In the third article information about 7 metabolites found for GLP-1 was annotated into database.

The fourth article presents a new approach that stores the information coming from external sources such as MEROPS database in a chemically aware database (i.e. WebMetabase). In this study we exported peptide cleavage data from the MEROPS database (version 11 1/09/2017) for all the available proteases. In total information about 18760 substrate peptides and 21804 metabolites were extracted. All extracted data was uploaded into WebMetabase. Moreover, in this research we used an experimental dataset to perform metabolites identification using LC-HRMS data for the peptide dataset including the six commercially available peptides and five of them were synthetic analogues for the same peptide series, the luteinizing-hormone releasing hormone.

2.3 Third step

In the third step, we implemented an algorithm that performs frequency analysis (FA) based on the exact match of residue structures in the site of cleavage (so-called simple frequency analysis, SFA) (Figure 9). This approach can be used to discover the most frequent scissile bonds within the generated database depending on the protease based on the sequence of interest. SFA allows us to create a set of empirically derived rules that define protease specificity rules that later can be used to predict the metabolic liability of different amide bonds in a new non-tested peptide. Consequently, to make the prediction one can do a search for the exact match of motifs in the peptide of interest to the cleavage sites considered by frequency analysis within the database. The highest prediction rate will be given to the found motifs with the highest frequency. In Figure 11 the Mass-MetaSite/WebMetabase workflow from experimental data to searchable information in a database manageable by *in silico* analysis tools is shown.

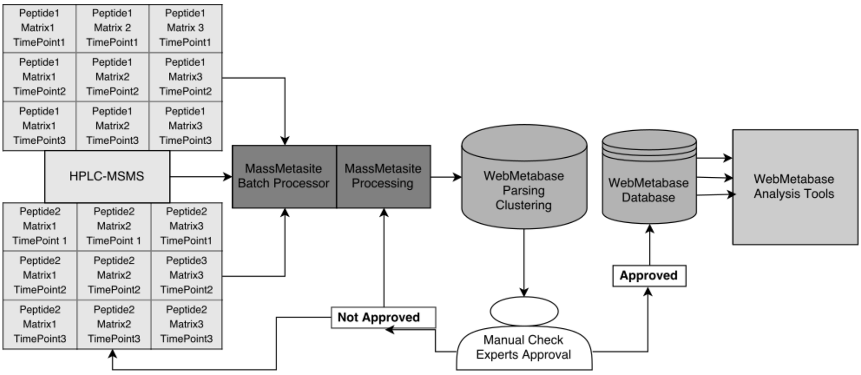

Peptide1 Matrix1 TimePoint1
Peptide1 Matrix 2 TimePoint1
Peptide1 Matrix 3 TimePoint1
Peptide1 Matrix1 TimePoint2
Peptide1 Matrix2 TimePoint2
Peptide1 Matrix3 TimePoint2
Peptide1 Matrix1 TimePoint3
Peptide1 Matrix2 TimePoint3
Peptide1 Matrix3 TimePoint3
HPLC-MSMS
Peptide2 Matrix1 TimePoint 1
Peptide2 Matrix2 TimePoint 1
Peptide2 Matrix3 TimePoint1
Peptide2 Matrix1 TimePoint2
Peptide2 Matrix2 TimePoint2
Peptide3 Matrix3 TimePoint2
Peptide2 Matrix1 TimePoint3
Peptide2 Matrix2 TimePoint3
Peptide2 Matrix3 TimePoint3
MassMetasite Batch Processor
MassMetasite Processing
WebMetabase Parsing Clustering
WebMetabase Database
WebMetabase Analysis Tools
Approved
Not Approved
Manual Check Experts Approval

Figure 11. Mass-MetaSite/WebMetabase workflow for experimental data. © Radchenko T, Brink A, Siegrist Y, Kochansky C, Bateman A, Fontaine F, et al. (2017) Software-aided approach to investigate peptide structure and metabolic susceptibility of amide bonds in peptide drugs based on high resolution mass spectrometry. PLoS ONE 12(11): e0186461.

In the first article 77 distinct cleavage sites were found during the metabolite identification study. The most frequent observed cleavage sites always agreed with those already reported in MEROPS and ENZYME databases [49, 51, 52]. In addition, during examination of the effect of small chemical changes in peptide structures for two groups described above we found that elastase cleaved the Ser-Tyr bond and trypsin cleaved the Arg-Pro bond, except when D-Ser(tBu) was positioned in buserelin and goserelin on the P2’ position, and when the C-terminal Pro was modified to Pro-NHNHCONH2 in goserelin instead of Pro-NHet. In GLP-1 analogues group our approach revealed that liraglutide and GLP-1 were cleaved at the Ala-Glu linkage. Exenatide was not cleaved at the same site due to the amino acid change in the parent sequence where Ala8 was modified to Gly, taspoglutide was cleaved despite the modification Ala8 to α-aminoisobutyric acid.

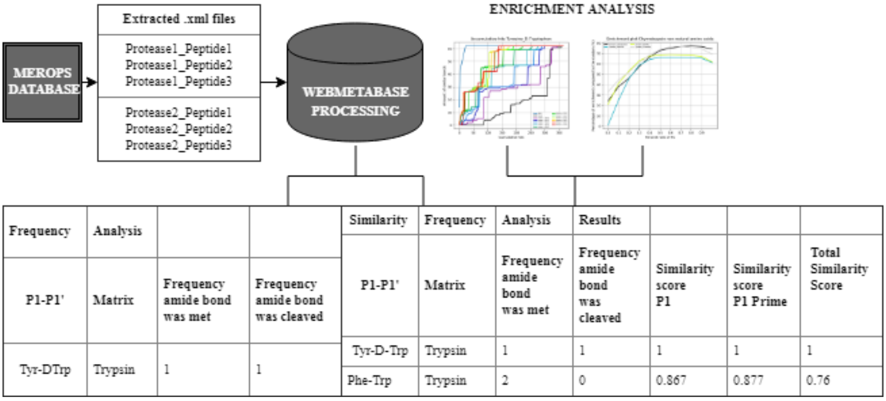
In the second article 8 distinct sites of cleavage were registered during frequency analysis.

In the third article 7 cleavage sites were found for GLP-1 that agreeded with cleavage sites reported in the literature with most of the rat metabolites being found by both unsupervised DIA methods [96].

The experimental results have been validated for those cases were literature data was available. The main advantages of the developed approach are the ability to store processed information in a searchable format within a database leading to frequency analysis of the labile sites for the analyzed peptide drugs. Also, in the fourth article, we have shown that the database can be enriched with new experimental data and subsequently customized for a peptide set of interest for further analysis. This new algorithm may be useful to optimize peptide drug properties with regards to cleavage sites, stability, metabolism and degradation products in drug discovery.

2.4 Fourth step

In the fourth step (Figure 9), we implemented an algorithm that performs similarity frequency analysis (SimFA). This analysis reveals the group of metabolically labile amide bonds defined by similar pharmacophoric or physicochemical properties towards the specific proteases/specific media. This methodology can be used to perform a frequency analysis to discover the most frequent cleavage sites for similar amide bonds, based on the similarity of the SB that participate in such a bond within the experimentally derived and/or public database. The site of cleavage is considered as similar to the SoC of interest when total similarity score is higher than a default total threshold score. Total similarity score is a linear combination of the similarity score for the SB in P1 position and the similarity score for the SB in P1’ position. Similarity score P1 describes similarity between pharmacophoric or physicochemical properties of SB in position P1 in the SoC of interest to the SB in P1 in the SoC of comparison. The workflow described above is represented in Figure 12.



Frequency	Analysis		
P1-P1'	Matrix	Frequency amide bond was met	Frequency amide bond was cleaved
Tyr-DTrp	Trypsin	1	1

Similarity	Frequency	Analysis	Results			
P1-P1'	Matrix	Frequency amide bond was met	Frequency amide bond was cleaved	Similarity score P1	Similarity score P1 Prime	Total Similarity Score
Tyr-D-Trp	Trypsin	1	1	1	1	1
Phe-Trp	Trypsin	2	0	0.867	0.877	0.76

Figure 12. MEROPS/WebMetabase workflow with similarity frequency analysis workflow. © Radchenko T, Fontaine F, Morettoni L, Zamora I. WebMetabase: cleavage sites analysis tool for natural and unnatural substrates from diverse data source (submitted)

In the fourth article we were able to compare the identified cleavage sites from both sources (experimental data and data from external sources). We performed a similarity frequency analysis for the selected SoCs from the combined dataset (MEROPS + experimental data) using chymotrypsin and trypsin1, while caspase 6, matrix metallopeptidase-2 were selected only from MEROPS. The SimFA results were analyzed to obtain empirically derived rules for the cleavage site appearance rules based on similarity of the pharmacophoric properties for the cleavage sites. To investigate the effect of the similarity to the P1 and P1’ structural blocks on the recovery of known metabolized bonds we performed an enrichment analysis. After we plotted the enrichment for each of the selected bonds for each individual protease as a function of the weight of P1 similarity in the total enrichment score. Enrichment plots for all the selected proteases and bonds in both MEROPS and experimental datasets are shown in Figure 13. These plots were used to perform analysis of the enrichment versus the weighted factor for the contribution of similarity to P1 structural block.

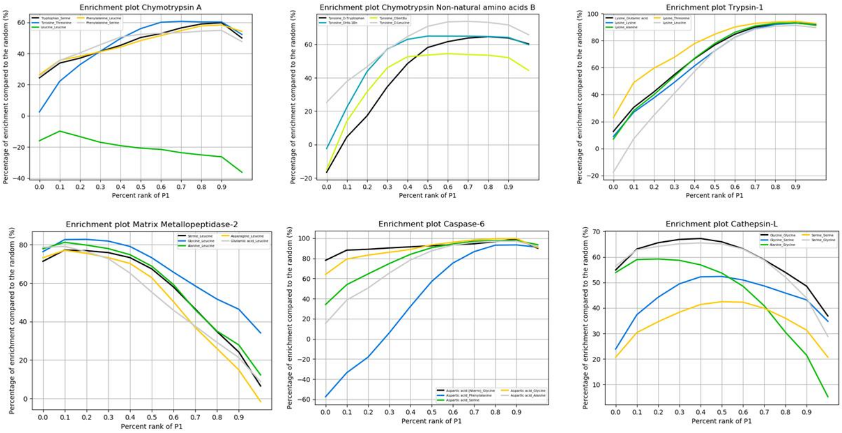


Figure 13. Enrichment plots for selected bonds from MEROPS dataset for all selected proteases and bonds with unnatural amino acids from experimental dataset for chymotrypsin. © Radchenko T, Fontaine F, Morettoni L, Zamora I. WebMetabase: cleavage sites analysis tool for natural and unnatural substrates from diverse data source (submitted)

In accordance with Figure 13 for chymotrypsin and trypsin1 the amino acids located on P1’ position in SoC have a minimal influence. Cconsidering only the similarity to the residue located in P1’ position, the percentage of enrichment is lower than 20% independently on SB on position P1’ used for the analysis. Meanwhile, when the contribution of P1 similarity to the total score is increased, the percentage of enrichment (PR) drastically increases. For example, for chymotrypsin it can be seen when Tyrosine, Tryptophan and Phenylalanine are the amino acids on position P1 and PR is reaching 90% that let us to collect information about the cleavage sites about 100% better than random. These results agree with literature [80] and in similarity matrix described in MEROPS database.

For caspase 6, the amino acids located on P1’ position in SoC has different influence depending on the amino acid (Figure 13). For example, on one hand when Phenylalanine is located on P1’ position and considering only the similarity to it the percentage of enrichment is worse than random. On the other hand, when Glycine is located on position P1’ and P1 is Asp the enrichment reaches the maximum comparing to random. These results agree with literature where it was described that caspase-6 cleavage preferentially occurs at sites composed of D|X where X is any amino acid but Pro, Glu, Asp, Gln, Lys, Arg [92, 97, 98].

For matrix metallopeptidase-2 (MMP-2), the amino acids located on P1 position in SoC has a minimal influence in the enrichment (Figure 13). Considering only the similarity to this amino acid in a cleavage site the percentage of enrichment is lower than 40% independently on SB on position P1 used for the analysis. Meanwhile, when the contribution of P1' increases the amount of collected information drastically increases. It is worth mentioning that in most of the cases Leu or Ile were located on position P1' in cleaved SoCs in MEROPS database. In accordance with literature MMP2 cleaves SoC where a Leu or other hydrophobic amino acid is in P1' position [99].

For cathepsin L, both amino acids located on P1 position and P1' position are important, but it also depends on the pair selected as a reference in the analysis (Figure 13). For example, if Gly is located on position P1' and Gly or Ser are in P1 position the influence of P1' is practically equals to P1 and maximum enrichment is reached when PR for P1 is 60%. Similarly, when Ser is located on position P1' and Gly or Ser are in P1 position their influence is equal. It means that maximum percentage of enrichment is reached when PRP1 and PRPr are equal to 0.5. These results agree with literature where it was described that preferentially cathepsin L cleaved at Gly in P1 and Gly or Ser in P1' position [100].

Moreover, we demonstrated that the algorithm can evaluate SoCs in peptide drugs from experimental dataset that contained nonstandard amino acids. In this case, we see that modification of the natural amino acid in position P1' to D-Ser-tBu, D-Leu or D-Trp does not improve the enrichment percentage and it is still close to zero or lower (random) when P1' influence is high. It does not change the specificity rules defined for the chymotrypsin based on MEROPS dataset. Meanwhile the contribution of P1 similarity increased the PR.

Since in the literature it was revealed that increasing the investigated local window's size around the cleavage site can improve the understanding of the protease specificity and probable proteolytic activity [87], we performed similarity frequency analysis for sites of cleavage with increased local window sequence P2-P2’ for caspase-6 and cathepsin L. Moreover, in the case of cathepsin L we also evaluated P4 - P4’. We concluded that influence of P1 and P1’ is not enough to be able to define protease specificity. We demonstrated that the influence of the P2 and P2’ depends on residues in the P1 and P1’ positions. Similar conclusion was reached in the case of cathepsin L for the P2 - P2’ and P4 - P4’ local windows.

In conclusion, a similarity frequency analysis of the actual SoC depending on the protease can be done to create a set of empirically derived rules based on molecular properties of the cleavage sites. These rules could be later used to predict the metabolic liability of different amide bonds in a new non-tested peptide. The proposed methodology as opposed to existing databases (i.e. ExPASy) can be applied in the case of non-natural amino acid and/or cyclic peptides. This approach can be used to derive cleavage site appearance rules for the specific peptide family (i.e. LHRH and analogues) or for specific experimental condition (i.e. individual protease or complex matrix as plasma). Moreover, since the system used to derive the cleavage site appearance rules (frequency analysis) could be linked to the software assisted metabolite structure elucidation based on MS data, the database is automatically enriched with the new experiments. Rules can then be refined to tune the system for the experimental conditions and/or peptide families of interest.

2.5 Fifth step

In the fifth step (Figure 9), we demonstrated that the implemented system and information stored in the created database can be used to train cleavage site prediction models using different machine learning algorithms based on molecular properties of the cleavage sites.

In the fifth article we used previously described database that contained data extracted from MEROPS and experimental data to train several models using different classifier learning approaches for eighteen proteases from four protease families: serine, cysteine, aspartic and matrix metalloproteases. Moreover, we developed models based on two different local window size P1-P1’ and P4-P4’. In the training dataset each sequence pattern around the potential cleavage site and actual site of cleavage was represented as a combination of Volsurf descriptors that characterized the physicochemical properties of the SBs in the sequence of the SoC. The described above workflow is shown in Figure 14.

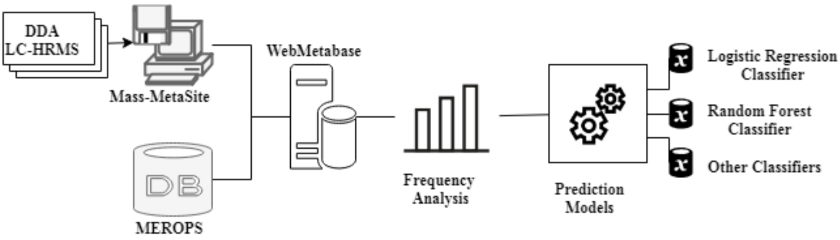


Figure 14. Cleavage sites prediction model’s preparation tool with MEROPS/WebMetabase workflow;

We compared the predictive performance of the models trained with different learning approaches applying 5-fold cross validation test and more importantly prediction results on an external dataset. Moreover, we examined the influence of the local window sequence size around the site of cleavage by comparing the models trained for P1-P1’ and P4-P4’ range. We revealed that the logistic regression and random forest classification models trained using window P4-P4’ outperformed other machine learning methods or the models trained using the P1-P1’ window.

We noted that training dataset size influenced the predictive performance of the models analyzing data for caspases. Finally, we compared the predictive performance of trained models with other approaches such as PROSPERous and SitePrediction tools. Logistic regression model recovered higher percentage of the known cleavage sites in the first 30% of the ranking positions comparing to the other tools. It can be explained by the fact that it performed better prediction on smaller peptides (Figure 15).

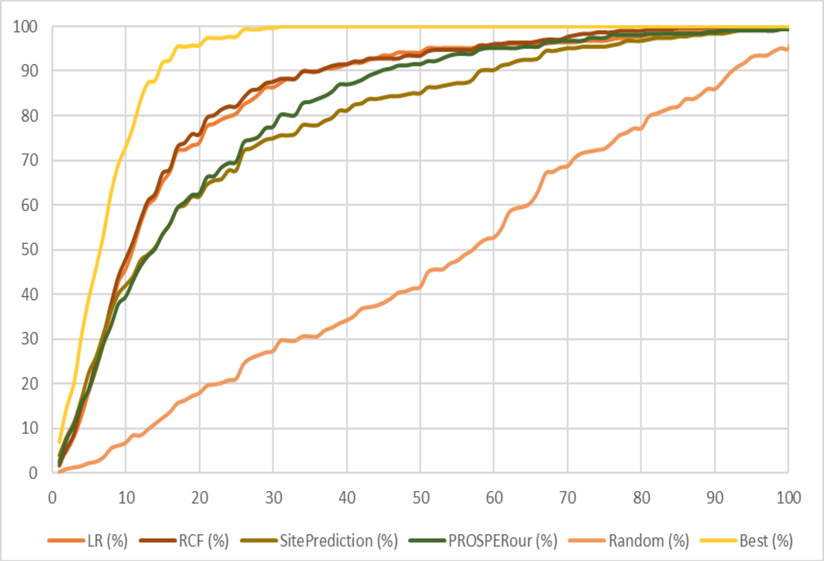


Figure 15. The cumulative ranking score in percentage for the recovered known sites of cleavage for selected protease families compared with PROSPERous and SitePrediction for 0-100% range of ranking positions percentage © Radchenko T, Fontaine F, Morettoni L, Zamora I. Software-aided workflow for predicting protease-specific cleavage sites using physicochemical properties of the natural and unnatural amino acids in peptide-based drug discovery (submitted).

Predicting possible sites of cleavage for individual proteases is an important task to be completed during drug-design process of peptide therapeutics to improve their stability and availability as a promising drug. In this study we presented a new approach in WebMetabase that helps to predict cleavage sites for the specific peptide family or for specific experimental condition (i.e. individual protease). One of the main advantages of this approach is that it generates a searchable database for the information coming from LC-MS based experimental data or from external sources such as MEROPS database. In this database each amino acid is described as a vector of physicochemical properties, Volsurf molecular descriptors, and/or pharmacophoric properties, SHOP descriptors. Thus, the motif around the potential cleavage was represented as combination of molecular descriptors. Comparing to MEROPS this database type can be enriched with new experimental or external data. This way to store the data can be utilized to perform frequency analysis to discover the most frequent scissile bonds within the generated database. The FA results can be used to derive a cleavage site appearance rules based on molecular properties of the cleavage sites and similarity of the molecular descriptors. These results can be used to train predictive models for individual proteases or complex matrices. We demonstrated that predictive performance of the models trained with different learning approaches can recover same or higher percentage of the known cleavage site in the first 30% of the ranking positions comparing to the other publicly available tools. The developed system can be linked to the software assisted metabolite structure elucidation based on MS data, the database is automatically enriched with the new experiments. Moreover, models can be re-trained with updated dataset and derived rules can be refined to fine tune the system for the experimental conditions and/or peptide families of interest. This knowledge can be applied during the design-make-test drug discovery cycle.

3. FUTURE WORK

The main goal of the future work is to improve the predictive performance of the models and following steps can be applied to achieve this objective:

- According to the literature secondary structure of the potential cleavage site can influence certain proteases activity [102, 103] and considering secondary structure, flexibility of the potential cleavage site region and solubility during models training can improve predictive performance of the models [87,101]. It was suggested by Song et al. that considerable unfolding for a helical segment to bind into the active sites of a protease in a manner appropriate for catalysis is required during protease-substrate interactions. Moreover, a solvent accessible surface is a key factor that determines whether a substrate can be accessed and cleaved by the protease [101]. In the present work we evaluated only predictive models trained on physicochemical properties of the residue containing in the cleavage site. Therefore, we planned to train predictive models based on pharmacophoric properties and add consideration of the secondary structure of the cleavage site and compare the predictive performance of the generated models for the state of the art prediction tools (e.g. PROSPERous).
- Since it was demonstrated by Song et al. that predictive models based on a bigger local window from P4-P2’ to P8-P8’ sites achieve the better performance depending on protease [87,91,101] we planned to use different symmetrical and asymmetrical local window sizes to perform extensive feature selection to extract more relevant features and to compare predictive performance of the generated models for the state of the art prediction tools.
- Moreover, previous studies revealed that application of the combination of predictive models can improve the predictive performance of the models. We plan to combine several learning algorithms and use prediction from combination of prediction models based on Volsurf and SHOP.

4. CONCLUSIONS

1. We developed an algorithm for processing LC-MS peptide incubation data able to process the experimental data of any peptide structures up to 4000 Da in molecular weight incubated in multiple media. It was effectively applied in multiple acquisition modes and persist process data in the chemically aware database.We developed an annotation system that enabled the use of processed data for multiple applications. It allows to perform chemical search (substructure search or similar structures search), frequency analysis (count the number of different amino acids in P1 and P1' positions in cleavage site) and prediction of the potential cleavage site for different proteases. This system can handle any type of amino acid and can be enriched with new data and any kind of searches can be performed on it.

2. We developed a frequency analysis algorithm which, applied to the annotation system described before, is an effective tool to reveal the protease specificity rules. These rules are based on molecular properties of the cleavage sites for individual proteases and not limited to natural amino acids. Moreover, defined rules can be refined to tune the system for the experimental conditions and/or peptide families of interest.

3. We demonstrated the possibility to generate site of cleavage predictive models with a prediction performance comparable with those of state of the art prediction tools, but without their limitations on the type of peptide structure and/or proteases.

4. REFERENCES

1. Fosgerau K, Hoffmann T. Peptide therapeutics: Current status and future directions. *Drug Discov Today*. 2015:122-128. doi: 10.1016/j.drudis.2014.10.003.
2. Craik DJ, Fairlie DP, Liras S, Price D. The Future of Peptide-based Drugs. *Chem Biol Drug Des*. 2013;81(Rational Design of Biologics and Peptides):136-147. doi:10.1111/cbdd.12055.
3. Di L. Strategic Approaches to Optimizing Peptide ADME Properties. 2014;17(1):134-143. doi:10.1208/s12248-014-9687-3.
4. Manning MC, Chou DK, Murphy BM, Payne RW, Katayama DS. Stability of protein pharmaceuticals: An update. *Pharm Res*. 2010;27(4):544-575. doi:10.1007/s11095-009-0045-6.
5. Kaspar AA, Reichert JM. Future directions for peptide therapeutics development. *Drug Discov Today*. 2013;18(17-18):807-817. doi: 10.1016/j.drudis.2013.05.011.
6. Uhlig T, Kyprianou T, Martinelli FG, et al. The emergence of peptides in the pharmaceutical business: From exploration to exploitation. *EuPA Open Proteomics*. 2014. doi: 10.1016/j.euprot.2014.05.003.
7. Vlieghe P, Lisowski V, Martinez J, Khrestchatisky M. Synthetic therapeutic peptides: science and market. *Drug Discov Today*. 2010;15(1/2):40-56. doi: 10.1016/j.drudis.2009.10.009.
8. Banting FG, Best CH, Collip JB, Campbell WR, Fletcher AA. Pancreatic extracts in the treatment of diabetes mellitus. *Indian J Med Res*. 2007;125(3). doi:10.1097/00005053-192401000-00013.
9. Lau JL, Dunn MK. Therapeutic peptides: Historical perspectives, current development trends, and future directions. *Bioorganic Med Chem*. 2018;26(10):2700-2707. doi: 10.1016/j.bmc.2017.06.052
10. Bononi FC, Luyt LG. Synthesis and Cell-Based Screening of One-Bead-One-Compound Peptide Libraries. In: *Methods in*

Molecular Biology (Clifton, N.J.). Vol 1248.; 2015:223-237. doi:10.1007/978-1-4939-2020-4_15.

11. Henninot A, Collins JC, Nuss JM. The Current State of Peptide Drug Discovery: Back to the Future? *J Med Chem*. 2018;61(4):1382-1414. doi: 10.1021/acs.jmedchem.7b00318
12. Diller DJ. The synergy between combinatorial chemistry and high-throughput screening. *Curr Opin Drug Discov Devel*. 2008;11(3):346-355. http://www.ncbi.nlm.nih.gov/pubmed/18428088.
13. Kennedy JP, Williams L, Bridges TM, Daniels RN, Weaver D, Lindsley CW. Application of combinatorial chemistry science on modern drug discovery. *J Comb Chem*. 2008;10(3):345-354. doi:10.1021/cc700187t.
14. Saveaunu A, Datta R, Zhang S, et al. Novel Somatostatin-Dopamine Chimeric Compound Demonstrates Superior Efficacy in Suppressing Growth Hormone Secretion from Human Acromegalic Tumors Partially Responsive to Current Somatostatin and Dopamine Therapies. Paper presented at: ENDO 2016; Boston, MA.
15. Finan B, Ma T, Ottaway N, et al. Unimolecular dual incretins maximize metabolic benefits in rodents, monkeys, and humans. *Sci Transl Med*. 2013;5(209). doi:10.1126/scitranslmed.3007218.
16. Konkar A, Suckow A, Hummer T, et al. MEDI4166: a PCSK9 Ab-GLP-1 fusion molecule that elicits robust antidiabetic and antihyperlipidemic effects in rodents and non-human primates. Paper presented at: European Association for the Study of Diabetes 2016; Munich, Germany.
17. Tsomaia N. Peptide therapeutics: Targeting the undruggable space. *Eur J Med Chem*. 2015; 94:459-470. doi:10.1016/j.ejmech.2015.01.014.
18. Woodley JF. Enzymatic barriers for GI peptide and protein delivery. *Crit Rev Ther Drug Carrier Syst*. 1994;11(2-3):61-95. http://www.ncbi.nlm.nih.gov/pubmed/7600588.
19. Calcott MJ, Ackerley DF. Genetic manipulation of non-ribosomal peptide synthetases to generate novel bioactive

peptide products. *Biotechnol Lett*. 2014;36(12):2407-2416. doi:10.1007/s10529-014-1642-y.

20. Walsh CT. Insights into the chemical logic and enzymatic machinery of NRPS assembly lines. *Nat Prod Rep*. 2016;33(2):127-135. doi:10.1039/c5np00035a.
21. Romanova EV.; Sweedler JV. Peptidomics for the discovery and characterization of neuropeptides and hormones. Trends Pharmacol. Sci. 2015, 36, 579−586.
22. Koh CY, Kini RM. From snake venom toxins to therapeutics - Cardiovascular examples. *Toxicon*. 2012;59(4):497-506. doi: 10.1016/j.toxicon.2011.03.017.
23. Miller LJ, Sexton PM, Dong M, Harikumar KG. The class B G-protein-coupled GLP-1 receptor: an important target for the treatment of type-2 diabetes mellitus. *Int J Obes Suppl*. 2014;4(S1): S9-S13. doi:10.1038/ijosup.2014.4.
24. McGivern JG. Ziconotide: A review of its pharmacology and use in the treatment of pain. *Neuropsychiatr Dis Treat*. 2007;3(1):69-85. doi:10.2147/nedt.2007.3.1.69.
25. Escano J, Smith L. Multipronged approach for engineering novel peptide analogues of existing lantibiotics. *Expert Opin Drug Discov*. 2015;10(8):857-870. doi:10.1517/17460441. 2015.1049527.
26. Ross RP. Bioengineering lantibiotics for therapeutic success. Front. Microbiol. 2015, 6, 1363.
27. Sandiford SK. Advances in the arsenal of tools available enabling the discovery of novel lantibiotics with therapeutic potential. *Expert Opin Drug Discov*. 2014:1-15. doi:10.1517/17460441.2014.877882.
28. Liskamp RMJ, Rijkers DTS, Kruijtzer JAW, Kemmink J. Peptides and Proteins as a Continuing Exciting Source of Inspiration for Peptidomimetics. *ChemBioChem*. 2011;12(11):1626-1653. doi:10.1002/cbic.201000717.
29. Proulx C, Sabatino D, Hopewell R, Spiegel J, Garcia Ramos Y, Lubell WD. Azapeptides and their therapeutic potential. *Future Med Chem*. 2011;3(9):1139-1164. doi:10.4155/fmc.11.74.

30. Rafi SB, Hearn BR, Vedantham P, Jacobson MP, Renslo AR. Predicting and improving the membrane permeability of peptidic small molecules. J Med Chem. 2012;55(7):3163–9.

31. Chin JW. Expanding and Reprogramming the Genetic Code of Cells and Animals. *Annu Rev Biochem*. 2014; 83:379-408. doi:10.1146/annurev-biochem-060713-035737.

32. Maini R, Umemoto S, Suga H. Ribosome-mediated synthesis of natural product-like peptides via cell-free translation. *Curr Opin Chem Biol*. 2016; 34:44-52. doi: 10.1016/j.cbpa.2016.06.006.

33. Malyshev DA, Romesberg FE. The expanded genetic alphabet. Angew. Chem., Int. Ed. 2015, 54, 11930−11944.

34. Chen T, Hongdilokkul N, Liu Z, Thirunavukarasu D, Romesberg FE. The expanding world of DNA and RNA. *Curr Opin Chem Biol*. 2016; 34:80-87. doi: 10.1016/j.cbpa. 2016.08.001.

35. Delgado C, Francis GE, Fisher D. The uses and properties of PEG-linked proteins. *Crit Rev Ther Drug Carrier Syst*. 1992;9(3-4):249-304. http://www.ncbi.nlm.nih.gov/pubmed/ 1458545.

36. Tibbitts J, Canter D, Graff R, Smith A, Khawli LA. Key factors influencing ADME properties of therapeutic proteins: A need for ADME characterization in drug discovery and development. *MAbs*. 2016;8(2):229-245. doi:10.1080/19420862. 2015.1115937.

37. Katsila T, Siskos AP, Tamvakopoulos C. Peptide and protein drugs: the study of their metabolism and catabolism by mass spectrometry. *Mass Spectrom Rev*. 2012;31(1):110-133. doi:10.1002/mas.20340.

38. Ewles M, Goodwin L. Bioanalytical approaches to analyzing peptides and proteins by LC-MS/MS. *Bioanalysis*. 2011;3(12):1379-1397. doi:10.4155/bio.11.112.

39. Last RL, Jones AD, Shachar-Hill Y. Towards the plant metabolome and beyond. *Nat Rev Mol Cell Biol*. 2007;8(2):167-174. doi:10.1038/nrm2098.

40. Nowatzke WL, Rogers K, Wells E, Bowsher RR, Ray C, Unger S. Unique challenges of providing bioanalytical support for

biological therapeutic pharmacokinetic programs. *Bioanalysis*. 2011;3(5):509-521. doi:10.4155/bio.11.2.

41. Stenberg P, Luthman K, Artursson P. Prediction of membrane permeability to peptides from calculated dynamic molecular surface properties. *Pharm Res*. 1999;16(2):205-212. http://www.ncbi.nlm.nih.gov/pubmed/10100304.
42. Akamatsu M, Fujikawa M, Nakao K, Shimizu R. In silico prediction of human oral absorption based on QSAR analyses of PAMPA permeability. *Chem Biodivers*. 2009;6(11):1845-1866. doi:10.1002/cbdv.200900112.
43. Di L, Feng B, Goosen TC, Lai Y, Steyn SJ, Varma MV, et al. A perspective on the prediction of drug pharmacokinetics and disposition in drug research and development. Drug Metab Dispos. 2013;41(12):1975–93. 2011;3(5):509–21.
44. Chen T, Mager DE, Kagan L. Interspecies modeling and prediction of human exenatide pharmacokinetics. *Pharm Res*. 2013;30(3):751-760. doi:10.1007/s11095-012-0917-z.
45. Puente XS, Velasco G, Gutiérrez-Fernández A, Bertranpetit J, King MC, López-Otín C. Comparative analysis of cancer genes in the human and chimpanzee genomes. *BMC Genomics*. 2006; 7:1-9. doi:10.1186/1471-2164-7-15.
46. Bayden AS, Gomez EF, Audie J, Chakravorty DK, Diller DJ. A Combined Cheminformatic and Bioinformatic Approach to Address the Proteolytic Stability Challenge in Peptide-Based Drug Discovery. 2015;104(6):775-789. doi:10.1002/bip.22711.
47. Niedermeyer THJ, Strohalm M. mMass as a Software Tool for the Annotation of Cyclic Peptide Tandem Mass Spectra. 2012;7(9). doi: 10.1371/journal.pone.0044913.
48. Puente XS, Sánchez LM, Overall CM, López-otín C. Human and mouse proteases: A comparative genomic approach. 2003;4(July). doi:10.1038/nrg1111.
49. Rawlings ND, Waller M, Barrett AJ, Bateman A. MEROPS: the database of proteolytic enzymes, their substrates and inhibitors. 2014;42(October 2013):503-509. doi:10.1093/nar/ gkt953.

50. Igarashi Y, Eroshkin A, Gramatikova S, et al. CutDB: A proteolytic event database. *Nucleic Acids Res*. 2007;35(SUPPL. 1):546-549. doi:10.1093/nar/gkl813

51. Bairoch A. The ENZYME database in 2000. *Nucleic Acids Res*. 2000;28(1):304-305. doi:10.1093/nar/28.1.304.

52. Rawlings ND, Barrett AJ, Thomas PD, Huang X, Bateman A, Finn RD. The MEROPS database of proteolytic enzymes, their substrates and inhibitors in 2017 and a comparison with peptidases in the PANTHER database. *Nucleic Acids Res*. 2018;46(D1): D624-D632. doi:10.1093/nar/gkx1134.

53. Dwivedi P, Schultz AJ, Hill HH. Metabolic Profiling of Human Blood by High Resolution Ion Mobility Mass Spectrometry (IM-MS). *Int J Mass Spectrom*. 2010; 298:78-90. doi:10.1016/ j. ijms.2010.02.007.

54. Wu H, Zhang X, Liao P, et al. NMR spectroscopic-based metabonomic investigation on the acute biochemical effects induced by Ce(NO3)3 in rats. *J Inorg Biochem*. 2005;99(11):2151-2160. doi: 10.1016/J.JINORGBIO.2005.07. 014.

55. Viant MR, Rosenblum ES, Tieerdema RS. NMR-based metabolomics: a powerful approach for characterizing the effects of environmental stressors on organism health. *Environ Sci Technol*. 2003;37(21):4982-4989. http://www.ncbi.nlm.nih. gov/ pubmed/ 14620827.

56. Goodwin CR, Fenn LS, Derewacz DK, Bachmann BO, McLean JA. Structural Mass Spectrometry: Rapid Methods for Separation and Analysis of Peptide Natural Products. *J Nat Prod*. 2012;75(1):48-53. doi:10.1021/np200457r.

57. Distler U, Kuharev J, Navarro P, Levin Y, Schild H, Tenzer S. Drift time-specific collision energies enable deep-coverage data-independent acquisition proteomics. *Nat Methods*. 2014;11(2):167-170. doi:10.1038/nmeth.2767.

58. Wysocki VH, Resing KA, Zhang Q, Cheng G. Mass spectrometry of peptides and proteins. *Methods*. 2005;35(3):211-222. doi: 10.1016/j.ymeth.2004.08.013.

59. Jeanne Dit Fouque K, Garabedian A, Porter J, et al. Fast and Effective Ion Mobility–Mass Spectrometry Separation of d -

Amino-Acid-Containing Peptides. *Anal Chem*. 2017;89(21):11787-11794. doi: 10.1021/acs.analchem.7b03401.

60. Cui Q, Lewis IA, Hegeman AD, et al. Metabolite identification via the Madison Metabolomics Consortium Database. Nat Biotechnol. 2008;26(2):162-164. doi:10.1038/nbt0208-162.
61. Patti GJ, Yanes O, Siuzdak G. Innovation: Metabolomics: the apogee of the omics trilogy. *Nat Rev Mol Cell Biol*. 2012;13(4):263-269. doi:10.1038/nrm3314.
62. Hopfgartner G, Tonoli D, Varesio E. High-resolution mass spectrometry for integrated qualitative and quantitative analysis of pharmaceuticals in biological matrices. *Anal Bioanal Chem*. 2012;402(8):2587-2596. doi:10.1007/s00216-011-5641-8.
63. Makarov A, Scigelova M. Coupling liquid chromatography to Orbitrap mass spectrometry. *J Chromatogr A*. 2010;1217(25):3938-3945. doi: 10.1016/j.chroma.2010.02.022.
64. Glauser G, Grund B, Gassner A-L, et al. Validation of the Mass-Extraction-Window for Quantitative Methods Using Liquid Chromatography High Resolution Mass Spectrometry. *Anal Chem*. 2016;88(6):3264-3271. doi: 10.1021/acs.analchem. 5b04689.
65. Hopfgartner G. Can MS fully exploit the benefits of fast chromatography? *Bioanalysis*. 2011;3(2):121-123. doi:10.4155/ bio.10.191.
66. Bateman KP, Castro-Perez J, Wrona M, et al. MSE with mass defect filtering for in vitro and in vivo metabolite identification. *Rapid Commun Mass Spectrom*. 2007;21(9):1485-1496. doi:10.1002/rcm.2996.
67. Lapthorn C, Pullen F, Chowdhry BZ. Ion mobility spectrometry-mass spectrometry (IMS-MS) of small molecules: Separating and assigning structures to ions. *Mass Spectrom Rev*. 2013;32(1):43-71. doi:10.1002/mas.21349.
68. Xu C, Ma B. Software for computational peptide identification from MS–MS data. *Drug Discov Today*. 2006;11(13-14):595-600. doi: 10.1016/j.drudis.2006.05.011.

69. Prakash C, Shaffer CL, Nedderman A. Analytical strategies for identifying drug metabolites. *Mass Spectrom Rev*. 2007;26(3):340-369. doi:10.1002/mas.20128.

70. Castro-Perez JM. Current and future trends in the application of HPLC-MS to metabolite-identification studies. *Drug Discov Today*. 2007;12(5-6):249-256. doi: 10.1016/J.DRUDIS.2007. 01.007.

71. Ahn J, Gillece-Castro B, Berger S. BiopharmaLynx: A New Bioinformatics Tool for Automat ed LC /MS Peptide Map ping Assignment. http://www.waters.com/webassets/cms/library/docs/720002754en.pdf. Application note.

72. Zelesky V, Schneider R, Janiszewski J, Zamora I, Ferguson J, Troutman M. Software automation tools for increased throughput metabolic soft-spot identification in early drug discovery. *Bioanalysis*. 2013;5(10):1165-1179. doi:10.4155/bio.13.89.

73. Pähler A, Brink A. Software aided approaches to structure-based metabolite identification in drug discovery and development. *Drug Discov Today Technol*. 2013;10(1): e207-e217. doi: 10.1016/j.ddtec.2012.12.001.

74. Bonn B, Leandersson C, Fontaine F, Zamora I. Enhanced metabolite identification with MS(E) and a semi-automated software for structural elucidation. *Rapid Commun Mass Spectrom*. 2010;24(21):3127-3138. doi:10.1002/rcm.4753.

75. Cece-Esencan EN, Fontaine F, Plasencia G, et al. Software-aided cytochrome P450 reaction phenotyping and kinetic analysis in early drug discovery. *Rapid Commun Mass Spectrom*. 2016;30(2):301-310. doi:10.1002/rcm.7429.

76. Zamora I, Fontaine F, Serra B, Plasencia G. High-throughput, computer assisted, specific MetID. A revolution for drug discovery. *Drug Discov Today Technol*. 2013;10(1): e199-e205. doi: 10.1016/j.ddtec.2012.10.015.

77. Ahlqvist M, Leandersson C, Hayes MA, Zamora I, Thompson RA. Software-aided structural elucidation in drug discovery. *Rapid Commun Mass Spectrom*. 2015;29(21):2083-2089. doi:10.1002/rcm.7364.

78. Brink A, Fontaine F, Marschmann M, et al. Post-acquisition analysis of untargeted accurate mass quadrupole time-of-flight MS ^{E} data for multiple collision-induced neutral losses and fragment ions of glutathione conjugates. *Rapid Commun Mass Spectrom*. 2014;28(24):2695-2703. doi:10.1002/rcm.7062.

79. López-Otín C, Bond JS. Proteases: Multifunctional Enzymes in Life and Disease. *J Biol Chem*. 2008;283(45):30433-30437. doi:10.1074/jbc.R800035200.

80. Keil B. Specificity of proteolysis. Springer-Verlag Berlin-Heidelberg-NewYork, (1992) pp.335.

81. Schechter I, Berger A. On the size of the active site in proteases. I. Papain. *Biochem Biophys Res Commun*. 1967;27(2):157-162. http://www.ncbi.nlm.nih.gov/pubmed/6035483.

82. Schechter I, Berger A. On the active site of proteases. 3. Mapping the active site of papain; specific peptide inhibitors of papain. *Biochem Biophys Res Commun*. 1968;32(5):898-902. http://www.ncbi.nlm.nih.gov/pubmed/5682314.

83. Klein J, Eales J, Zürbig P, Vlahou A, Mischak H, Stevens R. Proteasix: A tool for automated and large-scale prediction of proteases involved in naturally occurring peptide generation. *Proteomics*. 2013;13(7):1077-1082. doi:10.1002/pmic. 201200493.

84. Igarashi Y, Heureux E, Doctor KS, et al. PMAP: databases for analyzing proteolytic events and pathways. *Nucleic Acids Res*. 2009;37(suppl_1): D611-D618. doi:10.1093/nar/gkn683.

85. Wang J, Yadav V, Smart AL, Tajiri S, Basit AW. Toward Oral Delivery of Biopharmaceuticals: An Assessment of the Gastrointestinal Stability of 17 Peptide Drugs. *Mol Pharm*. 2015;12(3):966-973. doi:10.1021/mp500809f.

86. Verspurten J, Gevaert K, Declercq W, Vandenabeele P. SitePredicting the cleavage of proteinase substrates. *Trends Biochem Sci*. 2009;34(7):319-323. doi: 10.1016/j.tibs.2009.04. 001.

87. Song J, Li F, Leier A, et al. PROSPERous: high-throughput prediction of substrate cleavage sites for 90 proteases with improved accuracy. Hancock J, ed. *Bioinformatics*. 2018;34(4):684-687. doi:10.1093/bioinformatics/btx670.

88. Boyd SE, Garcia de la Banda M, Pike RN, Whisstock JC, Rudy GB. PoPS: a computational tool for modeling and predicting protease specificity. *Proceedings IEEE Comput Syst Bioinforma Conf*. 2004:372-381. http://www.ncbi.nlm.nih.gov/pubmed/16448030.
89. Gasteiger E, Hoogland C, Gattiker A, Duvaud S, Wilkins MR, Appel RD et al. Protein Identification and Analysis Tools on the ExPASy Server; In: John M. Walker, editors. The Proteomics Protocols Handbook. Humana Press; 2005. pp 571-607.
90. Gasteiger E, Gattiker A, Hoogland C, Ivanyi I, Appel RD, Bairoch A. ExPASy: The proteomics server for in-depth protein knowledge and analysis. *Nucleic Acids Res*. 2003;31(13):3784-3788. http://www.ncbi.nlm.nih.gov/pubmed/12824418.
91. Wang M, Zhao X-M, Tan H, Akutsu T, Whisstock JC, Song J. Cascleave 2.0, a new approach for predicting caspase and granzyme cleavage targets. *Bioinformatics*. 2014;30(1):71-80. doi:10.1093/bioinformatics/btt603.
92. Backes C, Kuentzer J, Lenhof H-P, Comtesse N, Meese E. GraBCas: a bioinformatics tool for score-based prediction of Caspase- and Granzyme B-cleavage sites in protein sequences. Nucleic Acids Res. 2005;33(Web Server issue): W208-13. doi:10.1093/nar/gki433.
93. Wee LJK, Tan TW, Ranganathan S. CASVM: web server for SVM-based prediction of caspase substrates cleavage sites. *Bioinformatics*. 2007;23(23):3241-3243. doi:10.1093/bioinformatics/btm334.
94. Piippo M, Lietzén N, Nevalainen OS, Salmi J, Nyman TA. Pripper: prediction of caspase cleavage sites from whole proteomes. *BMC Bioinformatics*. 2010; 11:320. doi:10.1186/1471-2105-11-320.
95. Garay-Malpartida HM, Occhiucci JM, Alves J, Belizario JE. CaSPredictor: a new computer-based tool for caspase substrate prediction. *Bioinformatics*. 2005;21: i169-i176. doi:10.1093/bioinformatics/bti1034.
96. Sharma R, McDonald TS, Eng H, et al. In Vitro Metabolism of the Glucagon-Like Peptide-1 (GLP-1)-Derived Metabolites GLP-1(9-36) amide and GLP-1(28-36) amide in Mouse and Human Hepatocytes. Drug Metab Dispos. 2013;41(12):2148-2157. doi:10.1124/dmd.113.054254.

97. Talanian R V, Quinlan C, Trautz S, et al. Substrate specificities of caspase family proteases. *J Biol Chem*. 1997;272(15):9677-9682. http://www.ncbi.nlm.nih.gov/pubmed/9092497.

98. Stennicke HR, Renatus M, Meldal M, Salvesen GS. Internally quenched fluorescent peptide substrates disclose the subsite preferences of human caspases 1, 3, 6, 7 and 8. *Biochem J*. 2000;350 Pt 2:563-568. http://www.ncbi.nlm.nih.gov/pubmed/10947972.

99. Chen EI, Kridel SJ, Howard EW, Li W, Godzik A, Smith JW. A Unique Substrate Recognition Profile for Matrix Metalloproteinase-2. *J Biol Chem*. 2002;277(6):4485-4491. doi:10.1074/jbc.M109469200.

100. Biniossek ML, Nägler DK, Becker-Pauly C, Schilling O. Proteomic Identification of Protease Cleavage Sites Characterizes Prime and Non-prime Specificity of Cysteine Cathepsins B, L, and S. J Proteome Res. 2011;10(12):5363-5373. doi:10.1021/pr200621z.

101. Song J, Tan H, Perry AJ, et al. PROSPER: An Integrated Feature-Based Tool for Predicting Protease Substrate Cleavage Sites. Srinivasan N, ed. PLoS One. 2012;7(11): e50300. doi: 10.1371/journal.pone.0050300.

102. Barkan DT, Hostetter DR, Mahrus S, et al. Prediction of protease substrates using sequence and structure features. *Bioinformatics*. 2010;26(14):1714-1722. doi:10.1093/bioinformatics/btq267.

103. Hubbard SJ. The structural aspects of limited proteolysis of native proteins. Biochim Biophys Acta. 1998;1382(2):191-206. http://www.ncbi.nlm.nih.gov/pubmed/9540791.

104. Ahlstrom M., Ridderström M, Luthman K, et al. Virtual Screening and Scaffold Hopping Based on GRID Molecular Interaction Fields. J. Chem. Inf. Model. 2005; 45, 1313-1323. doi:10.1021/CI049626P.

105. Fontaine F, Cross S, Plasencia G, Pastor M, Zamora I. SHOP: A Method For Structure-Based Fragment and Scaffold Hopping. *ChemMedChem*. 2009;4(3):427-439. doi:10.1002/cmdc.200800355.

106. Baroni M, Cruciani G, Sciabola S, Perruccio F, Mason JS. A Common Reference Framework for Analyzing/Comparing Proteins and Ligands. Fingerprints for Ligands And Proteins

(FLAP): Theory and Application. *J Chem Inf Model*. 2007;47(2):279-294. doi:10.1021/ci600253e.

107. Milletti F, Strochi L, Sforna G, et al. New and Original pKa Prediction Method using Grid Molecular Interaction Fields J. Chem. Inf. Model. 2007; 47, 2172—2181. doi:10.1021/ CI700018Y.

108. Bergmann R, Linusson A, Zamota I SHOP: Scaffold HOPping by GRID-Based Similarity Searches J. Med. Chem. 2007; 50, 2708–2717, doi:10.1021/JM061259G.

109. Cruciani G, Pastor M, Guba W. VolSurf: a new tool for the pharmacokinetic optimization of lead compounds. *Eur J Pharm Sci*. 2000;11 S2:S29-39. http://www.ncbi.nlm.nih.gov/ pubmed/11033425.

110. Cruciani G, Crivori P, Carrupt P-A, Testa B. Molecular fields in quantitative structure–permeation relationships: the VolSurf approach. *J Mol Struct THEOCHEM*. 2000;503(1-2):17-30. doi:10.1016/S0166-1280(99)00360-7.

5. PUBLICATIONS

PUBLICATION 1

Software-aided approach to investigate peptide structure and metabolic susceptibility of amide bonds in peptide drugs based on high resolution mass spectrometry

Tatiana Radchenko, Andreas Brink, Yves Siegrist, Christopher Kochansky, Alison Bateman, Fabien Fontaine, Luca Morettoni, Ismael Zamora

PLoS ONE 12(11): e0186461. https://doi.org/10.1371/journal.pone.0186461

Radchenko T, Brink A, Siegrist Y, Kochansky C, Bateman A, Fontaine F, et al. Software-aided approach to investigate peptide structure and metabolic susceptibility of amide bonds in peptide drugs based on high resolution mass spectrometry. PLoS One. 2017 Nov 1;12(11):e0186461. DOI: 10.1371/journal.pone.0186461

PUBLICATION 2

Software assisted analysis for peptide drug metabolism

Tatiana Radchenko, Anna Escolà, Antoni Riera, Aurora Valeri,
Ismael Zamora

(Manuscript draft)

Software assisted analysis for peptide drug metabolism

Tatiana Radchenko (4,5), Anna Escolà (1,2), Antoni Riera (1,2), Aurora Valeri (3), Ismael Zamora (4,5)
1. Institute for Research in Biomedicine (IRB Barcelona), Barcelona, Spain;
2. Department de Química Inorgánica i Orgánica, Universitat de Barcelona, Barcelona, Spain;
3. Molecular Horizon S.R.L, Perugia, Italy;
4. Pompeu Fabra University, Barcelona Spain;
5. Lead Molecular Design, S.L, Sant Cugat del Vallés, Spain

Abstract

The interest in using peptide molecules as therapeutic agents is due to their high selectivity and efficacy. However, most peptide-derived drugs cannot be administered orally because of their instability in the gastrointestinal tract. To achieve better ADME properties the following chemical modifications are typically applied: substitution of the common L-amino acids to D-amino acids, cyclization of the peptide and others. Somatostatin or Somatotropin release-inhibiting factor (SRIF14) is a natural hormone that is being used as gastric anti-secretory drug as well as to treat growth hormone secretion disorders and endocrine tumors. The substitution of phenylalanine, using non-natural aromatic amino acids to enhance the aromatic interactions, naturally present in the hormone between Phe6, Phe7 and Phe11 has been studied before.

We used a new approach implemented in Mass-MetaSite and WebMetabase to process DDA LC-HRMS analytical data collected for a set of eight peptide drugs (somatostatin and seven synthetic analogue, containing non-standard amino acids) incubated with human serum. The effect of small chemical/monomer changes in the peptide structure with respect to the matrix catalyzed activity for serum was investigated in his peptide set. During the metabolite identification study in total 17 metabolites were found resulting in 8 distinct cleavage sites. We compared the percent of remaining parent peptide with respect to the time for all investigated peptides to compute the half life for each case. All compounds from the dataset were hydrolyzed with the different velocity. The most stable

compound was the one where Phe7 was replaced to Msa7, Trp8 to D-Trp8 and both Cys3 and Cys14 were replaced to D-Cys. It was digested significantly slower than somatostatin.

Introduction

Peptide drugs are well-suited for treatment in a wide range of therapeutic areas, such as diabetics, cancer, osteoporosis, hormone therapy, cardiovascular diseases and many more [1]. Today, more than 70 peptides are represented on the worldwide drug market [2,3]. Most of the peptide drugs are represented by native peptide analogues since their absorption, distribution, metabolism and elimination (ADME) properties, safety and toxicity profiles are known and easier to predict. Since natural peptides exist as a part of the natural pathways they are produced as an answer to the biological signal, processed, released, perform their function and then are rapidly metabolized to turn off the signal. Therefore, natural peptides have in general a short half-life. Fast extraction happens through proteolysis, and pH dependent hydrolysis in blood, gastro-intestinal (GI) tract, and liver with consequent renal filtration [1,2,4,5].

In case that the peptide drug is administered orally there are several enzymatic barriers that should be crossed to become a successful drug. It is well known that numerous human proteases are involved in peptide degradation. The most important barrier after oral administration is the lumen of the small intestine, which contains peptidases secreted from the pancreas (e.g. chymotrypsin), as well as cellular peptidases from mucosal cells. The second one would be the brush border membrane of the epithelial cells, which contains at least 15 different peptidases [6]. Therefore, several structure-based peptide design methodologies described below were developed to improve the stability of the peptide drug.

Many efforts have been done on developing of new techniques to improve the following peptide candidate drug properties: selectivity, solubility, stability, bioavailability, safety and toxicity [7]. The main efforts are spent on optimizing its ADME properties, increasing the half-life and/or improving the stability and selectivity using synthetic chemical modifications. The following modification techniques are typically implemented: introduce a limitation in the enzymatic degradation of the peptide through the identification of the possible cleavage sites followed by substitution of identified residues and/or protection against proteolytic degradation through enhancement on the secondary structure (e.g. insertion of a structure inducing probe (SIP)-tail, lactam bridges, stapling or clipping of peptide sequences or cyclization) [2,5]. These changes are applied during the design-make-test drug discovery cycle, with hopes of improving the physicochemical and pharmacokinetics properties of the compound of interest. Synthetic and modified peptides require more attention to the analysis of ADME and toxicity (ADMET) properties since the chemical structures of some of the canonical monomers used for the synthesis are modified and it is necessary to evaluate the potential ADMET properties of the new molecule produced.

Somatostatin or Somatotropin release-inhibiting factor (SRIF14) is a natural hormone that is being used as gastric anti-secretory drug as well as to treat growth hormone secretion disorders and endocrine tumors. The substitution of phenylalanine, using non-natural aromatic amino acids to enhance the aromatic interactions, naturally present in the hormone between Phe6, Phe7 and Phe11 has been studied before [8]. Octreotide was discovered as synthetic analogue of the natural hormone somatostatin.

We used a new approach implemented in Mass-MetaSite to process data-dependent acquisition (DDA) liquid chromatography high-resolution mass spectrometry (LC-HRMS) analytical data. This data was collected for a set of eight peptide drugs (somatostatin and seven synthetic analogues, containing non-standard amino acids) incubated with human serum. The samples obtained were used to perform metabolite identification, to reveal potential cleavage sites and to store the processed information in a searchable format within a database (WMB). During the metabolite identification study in total 17 metabolites were found resulting in 8 distinct cleavage sites. We compared the percent of remaining parent peptide with respect to the time for all investigated peptides to compute the half life for each case. Moreover, we evaluated the influence of the chemical modifications on the half-life time of the investigated compounds comparing to the value obtained for somatostatin. All compounds from the dataset were hydrolyzed with the different velocity. More stable compounds were the compounds where following replacements were done: both Phe7 to Msa7, Trp8 to D-Trp8 and/or both Cys3 and Cys14 to D-Cys.

We demonstrated that the developed approach can elucidate metabolite structure of cyclic peptides and those containing unnatural amino acids. The processed information obtained could be stored in a searchable format within a database leading to frequency analysis of the labile sites for the analyzed peptides. This new algorithm may be useful to optimize peptide drug properties with regards to cleavage sites, stability, metabolism and degradation products in drug discovery.

Materials and methods

Dataset

A metabolite identification study was performed using a peptide set that included eight compounds: one commercially available, somatostatin and seven synthetic analogues (Table 1). All eight peptides were cyclic and seven of them had unnatural amino acids. These peptides were used to investigate the effect of small chemical/monomer changes in the peptide structure with respect to the matrix catalyzed activity. All test compounds were dissolved in water to reach the concentration of a 6 mg/mL.

Table 1: Peptide-substrates structures and other characteristics

Name	Code	Molecular formula	Calculated exact mass
Somatostatin	-	C76H104N18O19S2	1636.7167
Msa7_DTrp8_DCys14_SRIF	05/006	C79H110N18O19S2	1678.7636
LOrn4_Msa7_DTrp8_SRIF	05/030	C78H108N18O19S2	1664.7480
Msa7_DTrp8_SRIF	05/031	C79H110N18O19S2	1678.7636
LOrn4_Msa7_DTrp8_DCys14_SRIF	05/035	C78H108N18O19S2	1664.7480
DAla1_DCys3,14_Msa7_DTrp8_SRIF	05/095	C79H110N18O19S2	1678.7636
DCys3,14_Msa7_DTrp8_SRIF**	03/064	C79H110N18O19S2	1678.7636

Incubations

The incubations were done using human serum (Sigma, H4522), that was slowly unfrozen and maintained at temperature 37°C. All incubations were conducted at 37°C. The detailed information regarding the incubation conditions is provided in Supplementary Table 1.

The final solution was dispensed into Eppendorf tube with adding 100 uL in each and then maintained at temperature 37ºC. Probe compounds were added to manually to have a final peptide concentration of 0.6 mg/mL. Each tube was used to investigate each time point, a total of eleven time points was studied (0, 5 min, 10 min, 30 min, 1 h, 2 h, 4 h, 8 h, 24 h, 30 h and 48 h) per compound. The incubation tubes were placed in heated shaker at 37°C and 1000 rpm shaking speed. All the reactions were started with the addition of the compound at the same time and quenched at an appropriate finished time. Incubation quenching was carried out by adding 400 uL of cold acetonitrile (the final probe concentration is 0.12 mg/mL) and the internal standard (labetalol) at a concentration of 0.6 uM. Following reaction quenching, the samples were refrigerated in acetone-carbon dioxide bath. The samples were then centrifuged in an Eppendorf 5810 R at 10000 rpm for 5 min at 4°C. 100 uL of resulting supernatant were transferred to clean glass vials and 0.5 uL was injected onto an Agilent Zorbax Eclipse Plus (C18 150 x 2.1 mm), 1.8 µm column via a Thermo Scientific Ultimate 3000 ultra-performance liquid chromatography (UPLC) autosampler. All time points were analyzed using a data-dependent MS/MS method. Negative control samples were prepared under the same conditions of the incubation (see above), containing serum but without adding probe compounds.

UPLC-MS/MS

The chromatographic separation of metabolites was performed using the Thermo Ultimate 3000 UPLC system. Agilent Zorbax Eclipse (C18 150 x 2.1 mm), 1.8 µm column was heated to 40°C. The mobile phase consisted of 0.1% TFA in water (eluent A) and 0.1% TFA in acetonitrile (eluent B) at a flow rate of 0.25 mL/min. The initial condition was 0% eluent B, which was increased via a linear gradient to 100% until 20 min. Eluent B was then ramped down to 0% until the end of the run at 22 min. Full MS scans were acquired in the Orbitrap mass over m/z 150-1700 range with resolution of 70000, with an automatic gain control (AGC) setting of 1e6 and maximum injection time of 120 ms. A “Top-5” method and peptide-specific inclusion lists were used for MS acquisition. Full scan MS/MS was a data dependent acquisition (DDA) using peptide-specific inclusion lists containing amide hydrolysis ions of multiple charge states ($z \geq 2$). These lists were generated using a

MOL file for each peptide and the software Mass-MetaSite 5.1.9. The five most intense peaks with charge state ≥ 2 were fragmented in the HCD collision cell with normalized collision energy of 30% and tandem mass spectrum was acquired in the Orbitrap mass analyzer with resolution of 17500, AGC of 5e5 and max injection time of 60. The DDA method settings employed a minimum AGC target of 8e3, intensity threshold of 1.3e5, chromatographic peak width of 4 s with an apex trigger between 5 to 10 s, isotopes excluded, dynamic exclusion of 10 s.

Data processing

All data acquired from the LC-MS system were processed using Mass-MetaSite 5.1.9 [10-13]. The Mass-MetaSite Batch Processor was used to process data without supervision. The produced output was automatically uploaded into the web application "WebMetabase 3.2.9" (Molecular Discovery Ltd, Middlesex, UK) [9, 14], where all the samples from the same experiment were clustered together for further analysis and interpretation. In WebMetabase (WMB) the detected chromatographic peaks were displayed together with the structural elucidation data for parent and metabolites [9].

The Mass-MetaSite settings used for the Mass-MetaSite Batch Processor are reported in Supplementary Table 2. The sample list used for the batch was generated in WMB mirroring the experimental design (enzymes, time points and instrument) and defined as a WebMetabase protocol. Settings used for the WMB protocol are given in Supplementary Table 3.

Mass-MetaSite

The application of Mass-MetaSite for the interpretation of small molecules metabolic stability data has been described previously [10-13]. Mass-MetaSite uses as inputs the 2D structure of the compound together with control and treated sample data files. The data can be processed sample-by-sample manually or in a batch mode with an automatic processing set. The data processing consists of two steps: on a first step automatic detection of the chromatographic peaks related to the parent and metabolites and on a second step structure elucidation of the potential metabolites based on the fragmentation pattern for each peak detected. Once the list of potential chromatographic peaks has been selected, Mass-MetaSite compares the m/z associated with each peak to all the possible theoretical metabolites based on a list of included biotransformation reactions (the hydrolysis of amide bonds for peptides). Then it generates all possible metabolites based on a predefined list of metabolic biotransformation reactions.

The overall principle for the structural elucidation of metabolites is a comparison of fragment ions obtained from the parent and the ones from the metabolites and then identifying mass shifts corresponding to the mass of the metabolite or common neutral losses. In addition to the above comparative fragmentation analysis, the fragmentation of the metabolite without comparison to the parent molecule is considered [9]. For each detected metabolite a score is assigned based on the number of matches/mismatches between the theoretical fragment m/z value and the m/z value observed in the MSMS spectrum as it has been described for small molecules [13]. Finally, Mass-MetaSite results are uploaded into the WebMetabase and manually checked and approved by the expert.

WebMetabase

All WebMetabase experimental settings are reported in Supplementary Table 3. Each experiment consisted of a set of samples, i.e. one sample per incubation time point. Mass-MetaSite processes every sample file as a separate unit. For each sample it collected the following information: metabolic scheme, structural fragment assignment, retention time, MS area, MS relative area and ppm for each structure. After, WebMetabase consolidates all these data in a single interpretation for the entire experiment and analyzes which metabolite peaks from each sample can be grouped based on the retention time and m/z. These consolidated substrates and metabolites are used for the next step in the analysis.

A new algorithm was introduced into WebMetabase to store the information about peptide in a chemically aware searchable format, including a system to perform searches based on matches of chemical substructure [9].

Results and Discussion

In this section, we present the results of applying our approach for the analysis of the investigated dataset. Firstly, it uses the peptide mode from Mass-MetaSite to process the DDA HRMS data and find chromatographic peaks related to the parent and then elucidate the metabolites structure. Secondly, these results were uploaded to WebMetabase followed by consolidation of all these data. This consolidated data was used for further analysis, i.e. evaluation of the identified cleavage sites and kinetic curves for the substrate and metabolites.

Metabolite identification was performed on the eight investigated compounds that were incubated with human serum. The compounds were structurally diverse, containing natural and unnatural amino acids (Table 1). The result of the peak detection step for the two main metabolites, using the Mass-MetaSite algorithm for somatostatin, and for 30-minute time sample shown in Figure 1. The metabolites listed in the figure are named by a shift in m/z (such as -71 or -128) with respect to the parent. The computed m/z values of the identified metabolites agreed with the predicted values. The metabolites have almost the same retention time of the parent and therefore one could even think that they could be a fragment of the parent. We know this is not the case, not only because the Mass-MetaSite algorithm, but also, we can see how the parent disappear and the metabolites appear (Figure 2), therefore one cannot be fragment of the other, since the proportion of both ions under the same fragmentation conditions should be very similar if one would be an in-source fragmentation of the other one.

Figure 1. Extracted ion chromatogram of somatostatin after 30 minutes of incubation.
Blue peak - parent peptide compound;
Green peaks - first generation of metabolites;
Marine peak - internal standard;

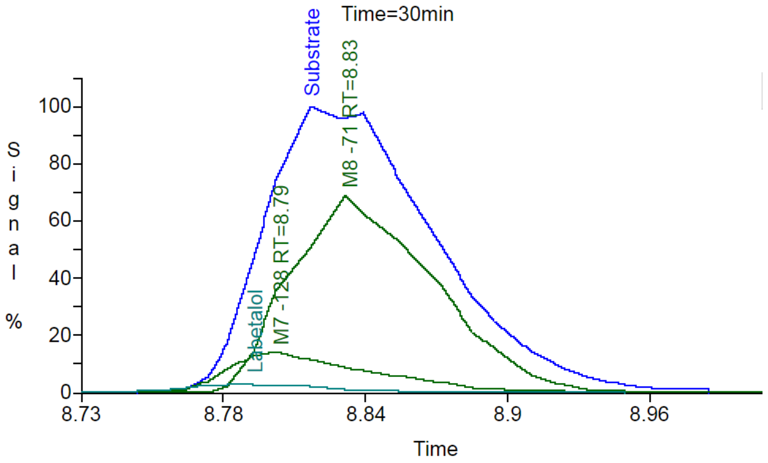
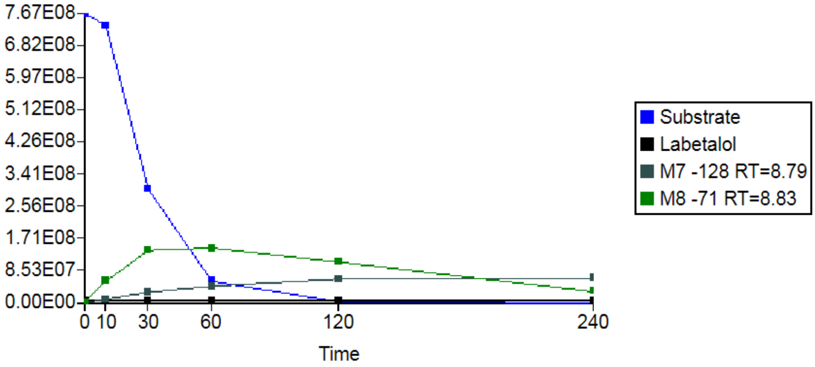
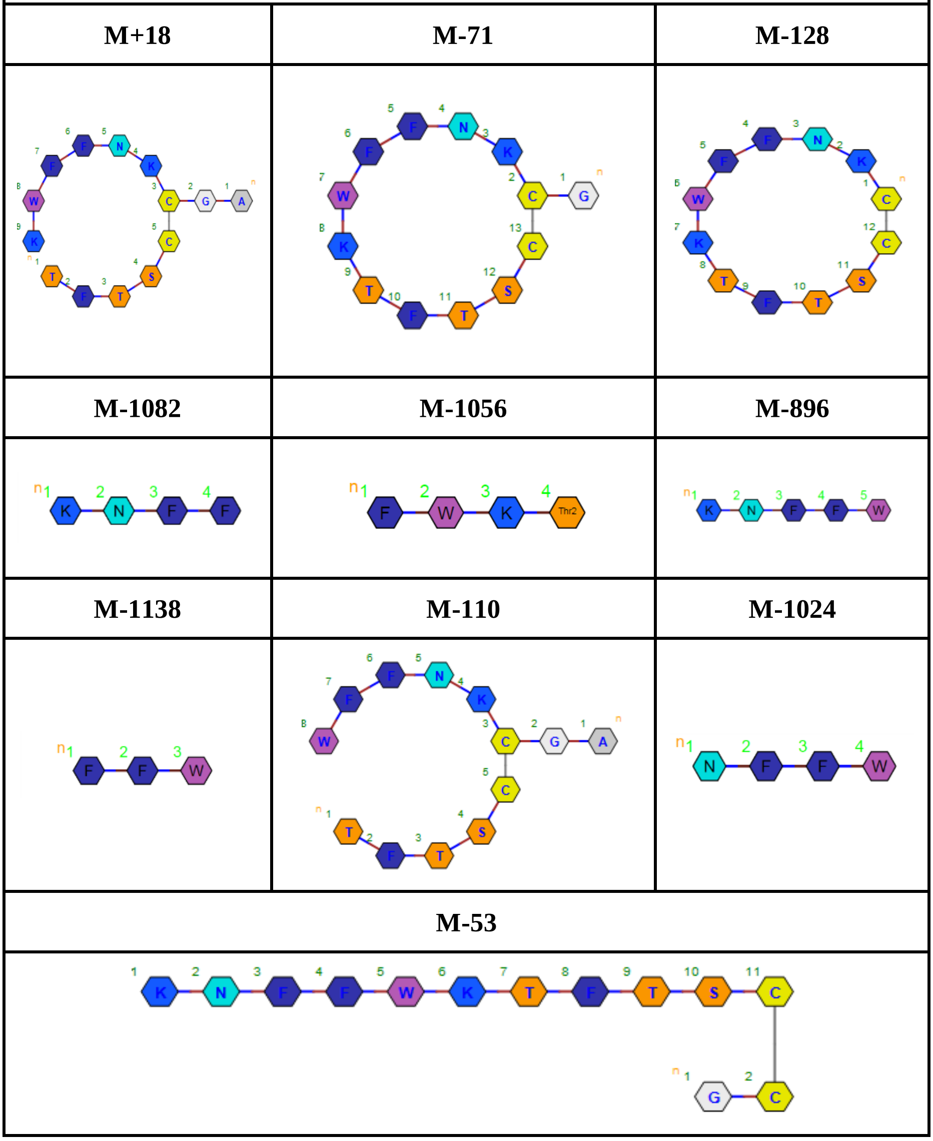


Figure 2. Appearance of the main metabolites in incubations with human serum and the Internal standard area at different time points and ratio of somatostatin



The second step of the algorithm assigns the chemical structures to the chromatographic peaks found for the metabolites. All the somatostatin and its analogues metabolites have similar fragmentation pattern compared to the substrate fragmentation. The metabolite fragment ions can have the same mass as a parent fragment or an expected mass shift, conserved and shifted, respectively. All assigned structures of the previously found metabolites for somatostatin and analogues are presented in Figure 3. All metabolites are presented in Supplementary Table 4 with the list of their substrates.

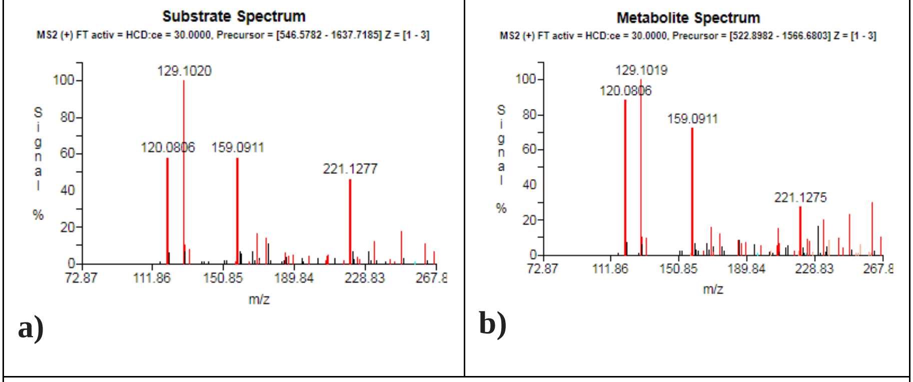
Figure 3. Proposed metabolites of somatostatin in incubations with human serum.



The structural assignments for somatostatin metabolites is shown in greater detail in Figure 4 and it is based on the fragment ions (ppm<10) that were detected in the substrate and metabolite spectra. In this figure, four main fragment ions were shown that are compatible with the structure for somatostatin. A score based on matching and mismatching fragments of the parent and metabolite is calculated and reported for each metabolite. For the metabolite M-71, the score was 2411 with 235 matching fragments and 27 mismatching fragments, respectively. The highest is the score the more confident the structural assignment is, although the score is an absolute value and therefore, it is difficult to compare from structure to structure, since there it will depend on the spectra obtained. In this case, the values obtained are considered high. In addition, the high proportion of matching peaks compared to the mismatches as well as the small difference between the observed and calculated m/z (<3 ppm) produce an additional confident in the proposed structure.

The analysis of the investigated peptides resulted in 17 metabolites that were annotated in the database. All the metabolites identified were produced by amide hydrolysis and were checked manually including a review of the assigned fragments. Metabolites were considered as reliable because the fragmentation was adequate, isotope pattern was as expected, the m/z small differences between the m/z of observed and theoretical, and the mass score was high. The metabolite time profile for the major metabolites of somatostatin and compound 095 is shown in Figure 5. Metabolite structure, Site of Cleavage (SoC) and the matrix were automatically registered into the database after experiment approval. The fact that all the information is stored in a consistent and chemically aware manner enables the classification of the predicted metabolites based on retention time and fragmentation across different experiments with different matrices and time points.

Fig 4. Full scan/data-dependent MS/MS fragment analysis of somatostatin substrate and metabolite M-.
a) Oxytocin substrate
b) Metabolite M-71
Red peaks - correlated with fragments that match between metabolite and fragment;
Orange peaks - correlated only with metabolite;
Blue peaks - correlated only with parent;



Substrate fragments and four highest matching fragments in metabolites

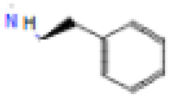
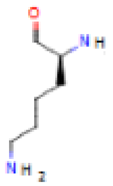
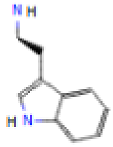
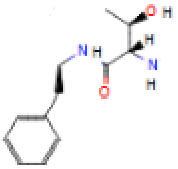
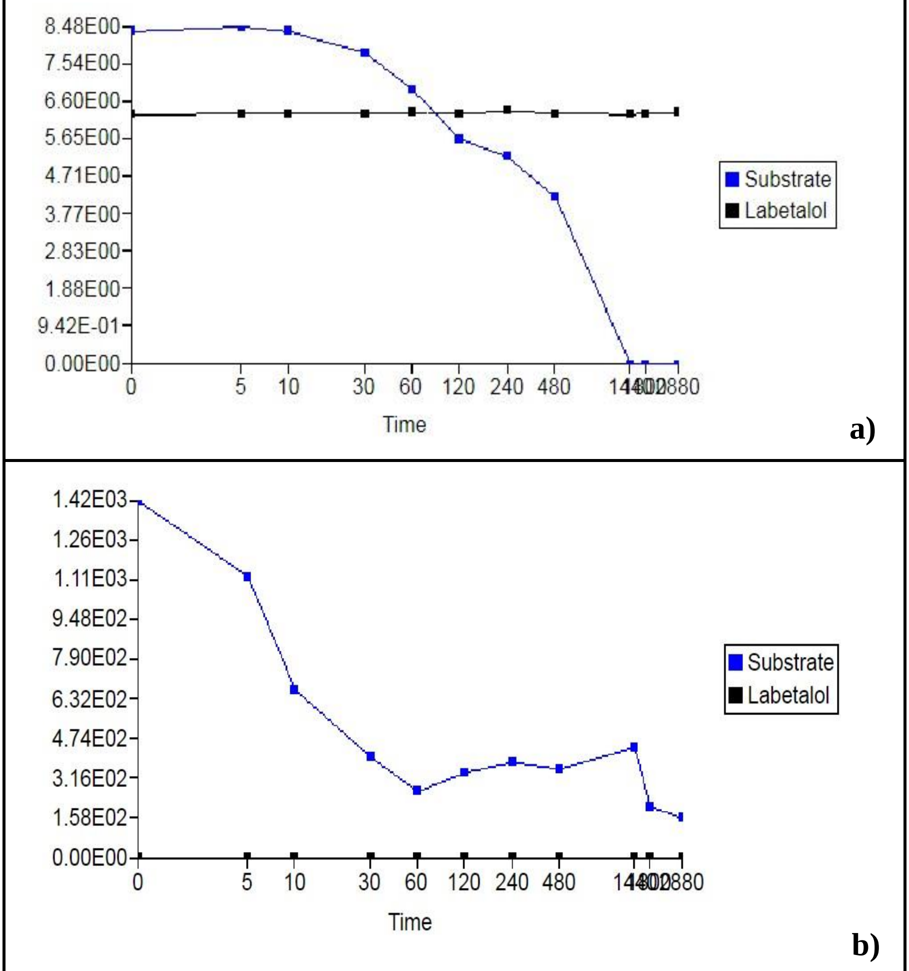
***m/z* 120.0806**	***m/z* 129.1020**	***m/z* 159.0911**	***m/z* 221.1275**
			

Figure 5. Metabolic stability of a) somatostatin and b) compound 095



To investigate injection-to-injection differences we used an internal standard during all incubations and so the peak area ratios of parent or metabolites to internal standard are shown in Figure 4. It is worth mentioning that the concentration of the metabolite cannot be directly correlated with the signal shown if a calibration line is not computed with an authentic standard of the metabolite. We did not have authentic standards of the metabolites and so these curves were evaluated qualitatively. The first generated metabolite usually has an exponential shape, when the metabolites are starting to be formed, for example M1-71 in Figure 6. If the metabolites are further metabolized, the signal of the metabolite will decrease since the metabolite has been consumed to generate a second generation one. Typically, the second-generation metabolite has a sigmoidal shape since it needs that the first-generation metabolite to form and then be further metabolized, for example being M2-230 in Figure 6. In some cases, metabolites could be detected in the sample labeled as t = 0, for example M5+18 in Figure 6. This is potentially because there was some insufficient time between the addition of quench reactants and it indicates the starting of the incubation at the moment of the extraction of the first sample.

For each investigated peptide half-life time was calculated. Firstly, a natural logarithm of determined area for substrate in each time point was computed and plotted against the incubation timeline. Then, a linear regression was applied to fit the line through several time points to describe substrate degradation and computed the R^2 to evaluate the linear regression model and select the best model. Finally, we calculated half-life time for each compound using slope of the line identified with linear regression. Results were shown in Table 2.

Figure 6. Appearance of the main metabolites in incubations with human serum and the Internal standard area at different time points and ratio of somatostatin and compound 095

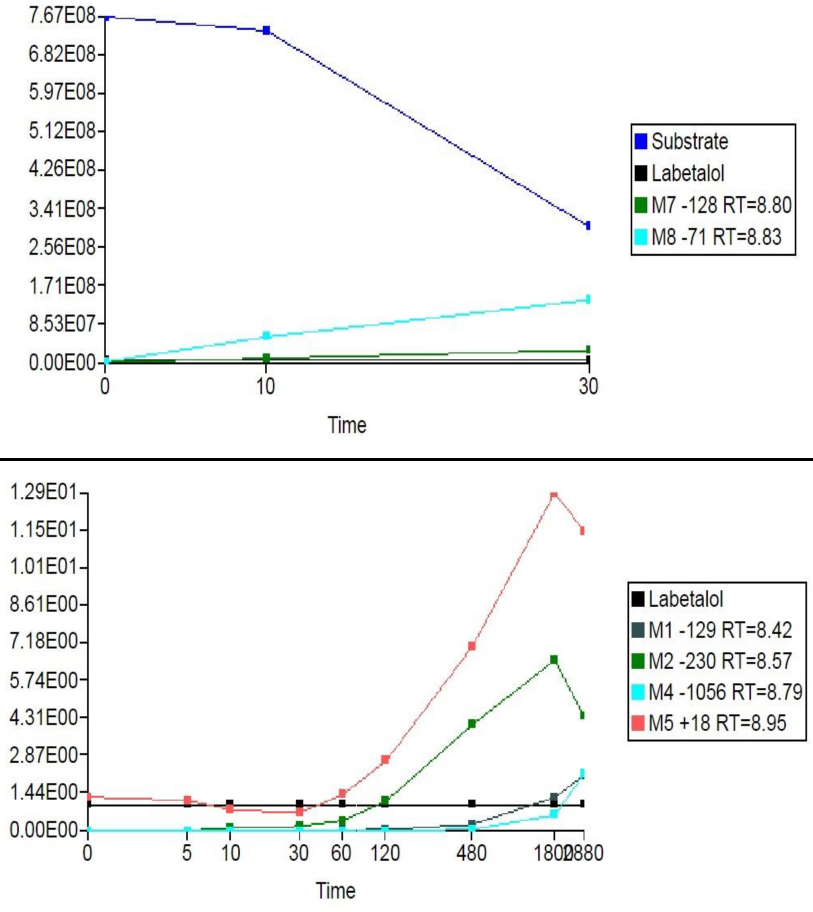


Table 2. Half-life time of compounds		
Compound name	**Half-life (slope)**	**R2 score**
Somatostatin	15 min	0.993
Compound 006	30 min	0.978
Compound 030	15 min	0.993
Compound 031	27 min	0.999
Compound 035	13 min	0.999
Compound 064	65 min	0.995
Compound 065	39 min	0.953
Compound 095	2400 min	0.957

All compounds from the dataset were hydrolyzed with the different velocity. We evaluated the effect of small chemical changes in peptide structures. The information about half-life time of the peptide compound and the structure of the firstly formed metabolite may help to understand the major metabolic clearance pathway and then aid in the designing of a new compound that would be more metabolically stable than the original compound. Compounds 006, 031, 064, 065 and 095 were more stable than somatostatin this may be due to the replacement of both Phe7 to Msa7 and Trp8 to D-Trp8. Compound 095 was digested significantly slower than somatostatin. This may be due to the replacement of both Cys3 and Cys14 to D-Cys. Our approach here revealed not only the rates of metabolism but also the site of catalysis. We found that all substrates except of 095 where cleaved in the linear part of the peptide.

Conclusions

In this research metabolite identification study was performed using a peptide set that included eight compounds: one commercially available, somatostatin and seven synthetic analogues. All test compounds were incubated in serum. All eight peptides were cyclic and seven of them had unnatural amino acids. We used a new approach implemented in Mass-MetaSite and WebMetabase to process experimental LC-HRMS data to find metabolite peaks, elucidate their structures, reveal potential cleavage sites. Moreover, we evaluated the effect of small chemical/monomer changes in the peptide structure with respect to the matrix catalyzed activity for serum. During this study in total 17 metabolites were found resulting in 8 distinct cleavage sites. We compared the percent of remaining parent peptide with respect to the time for all investigated peptides to compute the clearance for each case. All compounds from the dataset were hydrolyzed with the different velocity. Five compounds were more stable than somatostatin, this may be due to the replacement of both Phe7 to Msa7 and Trp8 to D-Trp8. Compound 095 was digested significantly slower than somatostatin and all other compounds due to the additional replacement of both Cys3 and Cys14 to D-Cys.

We demonstrated that the developed approach can elucidate metabolite structure of cyclic peptides and those containing unnatural amino acids, store processed information in a searchable format within a database. This new algorithm may be useful to optimize peptide drug properties with regards to cleavage sites, stability, metabolism and degradation products in drug discovery.

References

1. Di L. Strategic Approaches to Optimizing Peptide ADME Properties. AAPS J. 2015; 17:134. doi: 10.1208/s12248-014-9687-3.
2. Fosgerau K, Hoffmann T. Peptide therapeutics: Current status and future directions. Drug Discov Today. 2015;20(1):122-128. doi: 10.1016/j.drudis.2014.10.003.
3. Kaspar A, Reichert J. Future directions for peptide therapeutics development. Drug Discov Today. 2013;18(17-18):807-817. doi: 10.1016/j.drudis.2013.05.011.
4. Vlieghe P, Lisowski V, Martinez J, Khrestchatisky M. Synthetic therapeutic peptides: science and market. Drug Discov Today. 2010;15(1/2):40-56. doi: 10.1016/j.drudis.2009.10.009.
5. Tsomaia N. Peptide therapeutics: Targeting the undruggable space. Eur J Med Chem. 2015; 94:459-470. doi: 10.1016/j.ejmech.2015.01.014.

6. Woodley J. Enzymatic barriers for GI peptide and protein delivery. Crit. Rev. Ther. Drug Carrier Syst. (1994) 11, 61–95
7. Henninot A, Collins J, Nuss J. The Current State of Peptide Drug Discovery: Back to the Future? J. Med. Chem. 2018, 61, 1382−1414 doi: 10.1021/acs.jmedchem.7b00318
8. Martín-Gago P., Gómez-Caminals M, Ramón R, Verdaguer X, Martín-Malpartida P et al. Fine-tuning the π-π Aromatic Interactions in Peptides: New Somatostatin Analogs Containing Mesityl Alanine. Angew. Chem. Int. Ed., 2012, 51, 1820-1825.
9. Radchenko T, Brink A, Siegrist Y, Kochansky C, Bateman A, et al. Software-aided approach to investigate peptide structure and metabolic susceptibility of amide bonds in peptide drugs based on high resolution mass spectrometry; PlosOne, 2017.
10. Bonn B, Leandersson C, Fontaine F, Zamora I. Enhanced metabolite identification with MS(E) and a semi-automated software for structural elucidation. Rapid. Commun. Mass. Spectrom. 2010;24(21):3127-3138. doi: 10.1002/rcm.4753.
11. Cece-Esencan E, Fontaine F, Plasencia G, Teppner M, Brink A, et al. Software-aided cytochrome P450 reaction phenotyping and kinetic analysis in early drug discovery. Rapid. Commun. Mass. Spectrom. 2016;30(2):301-310. doi: 10.1002/rcm.7429.
12. Zamora I, Fontaine F, Serra B, Plasencia G. Metabolites: structure determination and prediction specific MetID. A revolution for drug discovery. Drug Discov. Today Technol. 2013;10(1): e199-e205. doi: 10.1016/j.ddtec.2012.10.015.
13. Zelesky V, Schneider R, Janiszewski J, Zamora I, Ferguson J, et al. Software automation tools for increased throughput metabolic soft-spot identification in early drug discovery. Bioanalysis. 2013;5(10):1165-1179. doi: 10.4155/bio.13.89.
14. Brink A, Fontaine F, Marschmann M, et al. Post-acquisition analysis of untargeted accurate mass quadrupole time-of-flight MS E data for multiple collision-induced neutral losses and fragment ions of glutathione conjugates. Rapid Commun Mass Spectrom. 2014;28(24):2695-2703. doi:10.1002/rcm.7062

Supplementary Table 1: Experiment conditions	
Matrix	Human Serum, Sigma 4522, pH (25°C) 7.4-7.6
Internal standard	Labetalol hydrochloride Sigma L1011
Mobile Phase A:	Water +0.1%FA
Mobile Phase B:	Acetonitrile+0.1%FA
Instrument	Thermo Scientific Q-Exactive
Injection Volume	0.5 uL

Supplementary Table 2: Mass-MetaSite settings are shown with experimental details		
	Mass-MetaSite Settings	
Import	Protonation policy	pH=7
	Maximum number of conformers	20
Metabolite generation	Minimum mass	50
	Metabolite stereochemistry and redundant metabolites	ignored
	MIM (the percentage of the monoisotopic mass of the parent)	30%
	Common cytochrome P450 reaction mechanisms	none
	Amide Hydrolysis	true
Mass settings, experiment	Retention time range (min)	not used
	GSH mode	deactivated
Mass settings, MS peaks	Maximum metabolite count limit	20
	Peak area threshold (%)	0.50%
	Peak area threshold (absolute)	0
	Peak detection smoothing	level 1
Expected metabolites	Rescue computed DRM peaks	not used
	Split computed DRM peaks	not used
	Adducts	not used
	Dimeric Ions	
	Unexpected metabolites	excluded
	Break metabolites	used
Mass settings, Met ID	Number of metabolite generations	2
	Compound fragmenting, bond breaking limit	2
	Even electron	MS and MS/MS
	Odd electron	MS and MS/MS
	N-Oxide	MS
Mass settings, DD-MS/MS algorithms, thresholds	Mass spectrometer	Thermo Q-Exactive_DDS
	Same peak tolerance (amu)	0.01
	Chromatogram automatic filtering threshold	0.97
	MS automatic filtering threshold	0.98
	MS/MS automatic filtering threshold	0.95
	Ionization mode	positive [M+H]$^+$
	Spectra comparisons for “Maximum MS/MS level"	2
	Signal filtering	automatic
	Scan filtering	automatic

Supplementary Table 3: WebMetabase settings with experimental details

User Experimental Settings	M/Z tolerance	0.025
	Retention time tolerance	0.2
	Retention time for calibration experiments	0.6
	Number for the important metabolites to show	4
Protocol Variables	Time	0, 5 min, 10 min, 30 min, 1h, 2h, 4h, 8h, 24h, 30h and 48h
	Matrix	Human serum

Supplementary Table 4: All identified metabolites with their substrates

Metabolite name	Substrates
M+18	Somatostatin, 006, 064, 095
M-71	Somatostatin, 006, 030, 031, 035, 064, 065
M-128	Somatostatin, 006, 030, 035, 065
M-110	Somatostatin, 006, 030, 035
M-1082	Somatostatin
M-1056	Somatostatin, 095
M-896	Somatostatin
M-1024	Somatostatin
M-1138	Somatostatin
M-129	095
M-762	065
M-53	Somatostatin, 030, 031, 035, 064
M-647	030
M-909	006, 064, 065, 095
M-230	065, 095
M-317	095
M-661	031

PUBLICATION 3

WebMetabase: cleavage sites analysis tool for natural and unnatural substrates from diverse data source

Tatiana Radchenko, Fabien Fontaine, Luca Morettoni, Ismael Zamora

(Manuscript submitted 06Apr2018, Bioinformatics journal)
(Revision sent 28Jun2018, Bioinformatics journal)

Radchenko T, Fontaine F, Morettoni L, Zamora I. WebMetabase: cleavage sites analysis tool for natural and unnatural substrates from diverse data source. Bioinformatics. 2018 Jul 25. DOI: 10.1093/bioinformatics/bty667

PUBLICATION 4

Software-aided workflow for predicting protease-specific cleavage sites using physicochemical properties of the natural and unnatural amino acids in peptide-based drug discovery.

Tatiana Radchenko, Fabien Fontaine, Luca Morettoni, Ismael Zamora

(Manuscript submitted 25May2018, PLoS One journal)

Radchenko T, Fontaine F, Morettoni L, Zamora I. Software-aided workflow for predicting protease-specific cleavage sites using physicochemical properties of the natural and unnatural amino acids in peptide-based drug discovery. PLoS One. 2019 Jan 8;14(1):e0199270. DOI: 10.1371/journal.pone.0199270

PUBLICATION 5

Metabolite Identification Using A Ion-Mobility Enhanced Data Independent Acquisition Strategy and Automated Data Processing.

Tatiana Radchenko; Christopher J Kochansky; Mark Cancilla; Mark Wrona; Russell J. Mortishire-Smith; Jayne Kirk; Gordon Murray; Fabien Fontaine; Ismael Zamora

(Manuscript draft)

Metabolite Identification Using A Ion-Mobility Enhanced Data Independent Acquisition Strategy and Automated Data Processing

Tatiana Radchenko[1,6]; Christopher J Kochansky[2]; Mark Cancilla[2]; Mark D Wrona[3]; Russell J. Mortishire-Smith[4]; Jayne Kirk[4]; Gordon Murray[5]; Fabien Fontaine[1]; Ismael Zamora[1,7]

[1]Lead Molecular Design, S.L., Sant Cugat Del Valles, Spain;

[2]Merck & Co., Inc., West Point, PA, USA;

[3]Waters Corporation, Milford, MA, USA;

[4]Waters Corporation, Wilmslow, United Kingdom;

[5]Waters Corporation, Beverly, MA, USA;

[6]Universitat Pompeu Fabra, Pl. de la Merce, 10-12, Barcelona, Spain;

[7]Molecular Discovery Ltd., London, United Kingdom;

Abstract:

Liquid chromatography/mass spectrometry (LC/MS) is an essential tool for efficient and reliable quantitative and qualitative analysis and underpins much of contemporary drug metabolism and pharmacokinetics. The characterization of small molecule clearance and metabolism using LC/MS is well understood and documented in the literature. Increasing attention on larger molecule therapeutics requires that optimized strategies for these kinds of chemotypes also be developed since there is the same requirement to optimize clearance, and for biotherapeutics containing non-native elements, to understand the metabolic fate of these components. Data-independent acquisition (DIA) methods such as MS^E (where E is collision energy) have reduced potential to miss metabolites, since product ion data are collected on all components, but do not formally generate quadrupole-resolved product ion spectra. Addition of ion mobility separation to DIA approaches such as High Definition Mass Spectrometry (HDMS^E) has the potential to both reduce the time needed to set up an experiment and maximize the chance that all metabolites present can be resolved and characterized. In this study, we report the comparison of DIA methods – MS^E and $HDMS^E$ using automated software processing with two commercially available software platforms, Mass-MetaSite and WebMetabase. We demonstrate that $HDMS^E$ is an effective approach for the elucidation of metabolite structures for small molecules and peptides, with excellent accuracy and quality. $HDMS^E$ provided high reproducibility and gave outcomes comparable with a state-of-the-art DDA workflow.

Introduction:

High performance liquid chromatography (HPLC) coupled to mass spectrometry (MS) is currently the method of choice for metabolite identification and quantification studies for small molecules, peptides and others [2-5]. Improvements in chromatographic separation such as ultra-high performance liquid chromatography (UPLC) have generally led to higher chromatographic resolution via sharper peaks and correspondingly higher MS sensitivity. A common strategy for MS data acquisition involves data-dependent acquisition (DDA) in which a list of likely metabolites is often employed to drive targeted fragmentation (MS^2). This methodology has the advantage that good quality MS^2 spectra are obtained for expected metabolites but has the possibility that unexpected metabolites will be missed. Furthermore, when the number of possible metabolites is high, the target list may be too large to be effectively used and may trigger MS^2 acquisition on isobaric background ions. Data-independent acquisition (DIA) methods such as MS^E (where E is collision energy) have reduced potential to miss metabolites, since product ion data are collected on all components, but do not formally generate quadrupole-resolved product ion spectra [10].

A variety of types of mass spectrometers can be coupled with liquid chromatography front ends, such as tandem quadrupole, quadrupole time-of-flight (QTOF) and Orbitrap mass spectrometers. Notwithstanding the sensitivity of these platforms, and the quality of the data which can be generated, their usage in drug metabolite identification can be a time-consuming task. An established DIA approach to data collection, such as ‘all-in-one’ fragmentation or MS^E [6], employs a rapid alternation between two full scan MS functions. The first scan function applies a low collision energy which results in precursor ion spectra (drug and metabolites), and the second scan function acquires data at high collision energy resulting in almost simultaneous acquisition of high resolution fragment ion spectra. The use of QTOF platforms with UPLC provides well-resolved peaks and in most cases the predominant fragment ions can be associated with a single matching precursor ion via peak convolution algorithms [6].

An extension of the MS^E approach was enabled by the introduction of ion mobility functionality into mass spectrometers (ion mobility-mass spectrometry, IM-MS). Briefly, IM-MS is a two-dimensional separation technique that separates ions in a dimension related to structure as a function of the ion's collision cross section (CCS) and subsequently in a second dimension according to the mass-to-charge ratio [7]. The CCS represents the area of the ion available for collisions with molecules in the gas phase. Many forms of ion mobility exist: high field asymmetric waveform ion mobility (FAIMS), differential ion mobility, traveling wave ion mobility (TWIMS) and uniform field ion mobility (IMS), although FAIMS is distinct in that it uses ion mobility to afford improved selectivity, and does not result in the generation of a CCS value. A specific feature of ion mobility spectroscopy, when coupled with mass spectrometry and a post-IMS collision cell, is that fragment ions can be correlated with their precursor ions on the basis of a shared drift time (the time taken to transit the ion mobility cell, subsequently converted to a CCS value) to generate IM-resolved spectra. This approach is particularly useful in complex samples [3] and high definition (HD) $HDMS^E$ approaches have been applied to a variety of biological, pharmaceutical and environmental scenarios [1]. IMS provides an additional dimension of separation improving chromatographic peak capacity while concomitantly reducing chimeric and composite interferences [7,8]. More importantly, the measurement of CCS as a consequence of IMS experiments affords a compound-specific parameter, which in many cases allows isobaric components to be discriminated [9]. A key aspect of the combination of an IMS separation (typically occurring in the millisecond time-frame) and MS detection (typically occurring in the microsecond time frame) is that it allows an additional separation step to be obtained on the MS time-scale (e.g., in addition to liquid chromatography), without compromising the speed of MS detection [9].

In this report we compare structural mass spectrometry techniques such as MS^E, $HDMS^E$ and DDA mass spectrometry approaches to peptide metabolite identification. While DDA is commonly used, it has a number of limitations, such as irreproducible precursor ion selection (in the case where only the most intense ions are selected, or the dynamic exclusion option is selected), undersampling and long instrument cycle times. Unbiased DIA strategies have been developed to overcome the limitations of DDA [8]. DIA approaches perform parallel fragmentation of multiple precursor ions, regardless of intensity or other characteristics, resulting in more complex but complete datasets. They enable the acquisition of a complete, unbiased sample record, enhancing quantification reproducibility. Using peak deconvolution algorithms, precursor ions can be correlated with their corresponding fragment ions. Thus, in contrast to DDA-based methods, which are intrinsically limited by scan time, DIA methods are theoretically limited only by the peak capacity (the number of peaks which can be discriminated based on retention time, *m/z* and CCS differences). $HDMS^E$ approaches use drift time correlation between precursor and product ions to achieve selectivity, compared with the peak deconvolution approach used in the MS^E DIA approach. Both MS^E and $HDMS^E$ experiments alternate between a low and high CE state on alternate scans, allowing collection of precursor and fragment ion information for all species in an analysis without the sampling bias inherent to DDA where a specific *m/z* value must be isolated before fragmentation [8].

Technological advances in mass spectrometry (MS) such as accurate mass high resolution instrumentation have fundamentally changed the approach to systematic metabolite identification over the past decade [15]. The process of metabolite identification has become largely facilitated and partly automated by cheminformatics approaches such as Mass-MetaSite (MMS) [10], Metabolynx-XS [16], UNIFI [20] and MetabolitePilot [17] which are able to propose metabolite structures based on the combination of metabolite prediction and interrogation of analytical mass spectrometry data. Here we describe the use of a software tool, Mass-MetaSite and its associated web-enabled platform, WebMetabase, for processing

DIA (MS^E and HDMS^E) data. These tools process datasets from both small molecule and peptide metabolic stability experiments to determine the specific metabolic sites and metabolic cleavage sites and then store the results in a chemically aware database, where chemical structure- and substructure-based searches can be performed. We demonstrate further that HDMS^E enables the elucidation of metabolite structures for small molecules and peptides with excellent accuracy and quality as compared to MS^E, providing similar data accuracy and quality to published DDA work.

Materials and methods

Sample Generation and Preparation

Metabolite identification was performed for glucagon-like peptide-1 (7-37) and verapamil hydrochloride purchased from Sigma-Aldrich. Both compounds were incubated at 5 µM substrate concentration with 1 mL of rat hepatocytes at 1 million cells/mL cell density in a 48-well plate while shaking in an incubator at 37°C and 5% CO_2 atmosphere. Multiple time points were sampled and mixed with 2-volumes of acetonitrile by vortexing. Samples were spun in a centrifuge for 20 min at 10°C and 4000 rcf. The verapamil samples were then diluted with 4-parts water to 1-part quenched incubation. Samples were then frozen and stored at -70°C until analyses by LC-HRMS.

Data Acquisition

The MS$^{\mathbf{E}}$ and HDMS^E methods were conducted on a Waters ACQUITY UPLC with Vion IMS QTof mass spectrometer operated by UNIFI.

UPLC

GLP-1 verapamil and their transformation products were analyzed by reversed phase UPLC using an ACQUITY BEH C18 2.1x100 mm column set to 40^{o}C and 50^{o}C for verapamil and GLP-1 respectively. Chromatography was performed on an ACQUITY I-Class UPLC (Waters Corp. Milford, MA) system configured with a fixed loop injection system. Mobile phase A and B were 0.1%

formic acid (Thermo Fisher Scientific, Waltham, MA) in Milli-Q water (Millipore, Burlington, MA) and 0.1% formic acid in LC-MS grade acetonitrile (Thermo Fisher Scientific, Waltham, MA) respectively. The injection volume was set to 5 µL and 10 µL using partial loop injection mode for verapamil and GLP-1 respectively. Verapamil sample components were separated by the following gradient at 0.5 mL/min flow rate: Held at 5% B for 0.5 mins, increased to 85% B at 1.25 mins, increased to 50% B at 2.75 mins and finally increased to 95% B at 3.25mins. GLP-1 sample components were separated by the following gradient at 0.55 mL/min flow rate: held at 2% B for 1 min, increased to 60% B at 2.5 mins, and finally increased to 95% B at 4 mins.

MS^E and HDMS^E

A VION IMS QToF (Waters Corp. Milford, MA) equipped with an electrospray ionization interface was used in this analysis. Tuning parameters were optimized to achieve the maximum transmission of parent ions and to provide detailed fragmentation patterns. The following parameters were used: capillary voltage: 1 kV, desolvation temperature: 600^{o}C (verapamil)/500^{o}C (GLP-1), desolvation gas flow: 1000 L/hr, source temperature: 125^{o}C, low collision energy: 6 eV, high collision energy ramp: 20-45 eV (verapamil)/30-40 eV (GLP-1) and analyzer mode: sensitivity. Leucine enkephalin was used as the external lock mass for both mass and CCS correction. The scan range was from 50 – 1000 *m/z* at a scan speed of 0.1 s for verapamil and from 200 – 2000 m/z at a scan speed of 0.125 s for GLP-1.

Data Processing

All data acquired from the LC/MS system were processed using a prototype version of Mass-MetaSite 5.1.9 (Molecular Discovery Ltd, Middlesex, UK) able to read data from UNIFI 1.9.2 (Waters Corporation, Milford, USA) using the built-in Application Programming Interface (API). The produced output was manually uploaded into the web application WebMetabase 3.2.9 (Molecular Discovery Ltd, Middlesex, UK), where all samples from the same experiment were clustered together for further analysis and interpretation. WebMetabase was used to review the detected chromatographic peaks together with the structural elucidation data for parent and metabolites.

Mass-MetaSite

The application of Mass-MetaSite for the interpretation of small molecule and peptide metabolic stability data has been described previously [10-12]. Mass-MetaSite (MMS) uses as input the 2D structure of the compound together with control and treated sample data files. Settings used for the processing of GLP-1 and verapamil MS data in MMS are presented in Supporting Tables 1 and 2. The data processing workflow consists of two steps. Step - 1 consists of automatic detection of the chromatographic peaks related to the parent compound. Step - 2 consists of structure elucidation of the potential metabolites based on the fragmentation pattern for each detected peak. Once the list of potential chromatographic peaks has been selected (Step - 1), MMS compares the *m/z* associated with each peak with all possible theoretical metabolites which can be generated using a list of biotransformation reactions. In this study, the only transformation of interest selected for GLP-1 was the hydrolysis of amide bonds. For processing of the verapamil dataset, the *Hepatocytes* mode, which includes both phase - 1 oxidation and phase - 2 conjugation reactions, was used. Mass-MetaSite then generates all possible metabolites which can be formed on the basis of the rules sets used.

The overall principle for the structural elucidation of metabolites is a comparison of fragment ions obtained from the parent (assigned from the incubation time t = 0 sample) and those fragment ions from the metabolites (t = incubation time), and then identification of mass shifts corresponding to the mass change of the biotransformation, or common neutral losses. In addition to the above comparative fragmentation analysis, peptide data processing is executed using theoretical fragmentation of the metabolite without comparison to the parent molecule. This fragmentation strategy is most advantageous in the case of cyclic peptides, where the metabolite could be a linear peptide (amide hydrolysis results in ring opening), in which fragmentation is expected to be significantly different compared to that of the parent.

A score is assigned to each metabolite based on the number of matches/mismatches between the theoretical fragment *m/z* value and the *m/z* value observed in the MSMS spectrum. Once MMS results have been uploaded into WebMetabase, they can be reviewed and approved by the expert.

WebMetabase

Each experiment consisted of a set of samples, i.e. one sample per incubation time point per matrix. Mass-MetaSite processes each sample as a separate entity and thus generates three main pieces of information for each sample: metabolic scheme, spectrometry data (structural fragment assignment) and outcomes (retention time, MS area, MS relative area, CCS and ppm mass error) for each found component. WebMetabase then consolidates all these data from the individual files into a single interpretation for the entire experiment (time/matrix) and analyzes which metabolite peaks from each sample can be clustered based on the retention time and *m/z*. This consolidated data was used for further processing, i.e. evaluation of number of identified peaks, evaluations of structures, background noise cleaning, and determining the kinetics of the parent peptide and metabolites. After this manual data interpretation, review and approval of the experiments, parent and metabolite structures were stored in the database. This interpretation is used for the subsequent data review.

Results and Discussion.

In this section, we present the results of comparing DIA (MS^E and $HDMS^E$) experiments processed with Mass-MetaSite and WebMetabase.

Metabolite identification was performed on incubations of GLP-1 (7-37) with rat hepatocytes. Results were compared with published metabolism of GLP-1 (9-36) following incubation with mouse and human hepatocytes by Sharma et al. [13]. There was good agreement between the metabolites identified in these experiments and those found for GLP-1 (9-36) as shown in Table 1. In addition, the majority of the metabolites found in rat hepatocytes were identified using either DIA methods. The five most abundant metabolites of GLP-1 (7-37) were then selected for detailed analysis with respect to their fragmentation and structural assignment. These metabolites are marked in bold in Table 1. Identified verapamil metabolites were consistent with previous npublished work in suspensions of rat hepatocytes as well as with incubations in plated rat hepatocytes by Walles et al. [21].

Table 1. Metabolites of GLP-1 (7-37) amide and GLP-1 (9-36) [13] in rat and human hepatocytes, respectively

Parent/ Metabolite	Peptide sequence for GLP-1 (7-37)	Peptide sequence for GLP-1 (9-36)
Parent	HAEGTFTSDVSSYLEGQAAKEFIAWLVKGRG	EGTFTSDVSSYLEGQAAKEFIAWLVKGR
M-208 RT=2.20	**EGTFTSDVSSYLEGQAAKEFIAWLVKGRG**	**Parent (9-36)**
M-495 RT=2.17	**FTSDVSSYLEGQAAKEFIAWLVKGR (12-37)**	**FTSDVSSYLEGQAAKEFIAWLVKGR (12-36)**
M-642 RT=2.15	**TSDVSSYLEGQAAKEFIAWLVKGRG (13-37)**	**TSDVSSYLEGQAAKEFIAWLVKGR** (13-36)
M-830 RT=2.18	DVSSYLEGQAAKEFIAWLVKGR (15-37)	DVSSYLEGQAAKEFIAWLVKGR (15-36)
M-945 RT=2.10	**VSSYLEGQAAKEFIAWLVKGR (16-37)**	**VSSYLEGQAAKEFIAWLVKGR (16-36)**
M-1219 RT=2.07	**YLEGQAAKEFIAWLVKGRG (19-37)**	**YLEGQAAKEFIAWLVKGR (19-36)**
M-1382 RT=2.03	**LEGQAAKEFIAWLVKGR (20-37)**	**LEGQAAKEFIAWLVKGR (30-36)**
M-1495 RT=2.00	EGQAAKEFIAWLVKGR (21-37)	EGQAAKEFIAWLVKGR (21-36)
GLP-1 (18-36)	Not found	SYLEGQAAKEFIAWLVKGR
GLP-1 (27-36)	Not found	EFIAWLVKGR

The results of the first step of peak detection using the Mass-

MetaSite algorithm for GLP-1 and verapamil for the 120 min incubations are shown in **Figures 1 and 2**, respectively. The metabolites listed in the figure are named by a shift in *m/z* (such as -945 or +148) with respect to the parent. The experimental *m/z* values of the identified metabolites agreed with the predicted values, being with in less than 3-ppm. All selected metabolites for GLP-1 and two metabolites for verapamil correspond to first-generation products (from a single reaction, green peaks). The single brown peak is indicative that multiple enzymatic reactions (2 or more) are required to generate the observed *m/z*. Though detected at earlier time points, three of the verapamil metabolites were second generation metabolites.

Figure 1. Extracted ion chromatograms for GLP-1 metabolites after 120 minutes of incubation with rat hepatocytes: a) MS^{E}; b) HDMS^{E}. Blue peak - parent compound; Green peaks - first generation metabolites; Aquamarine peak - internal standard;

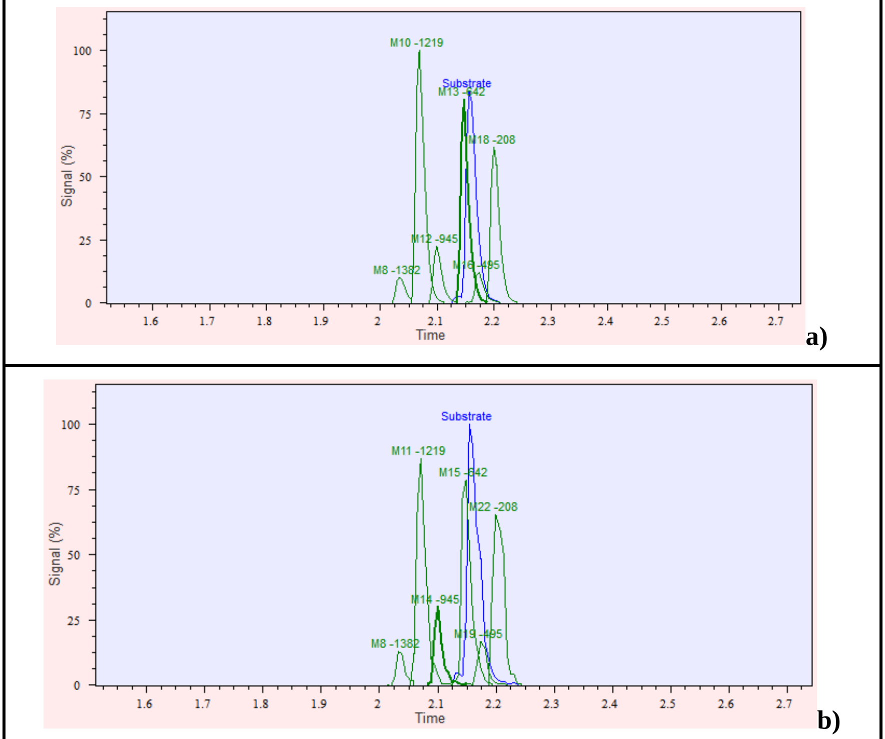
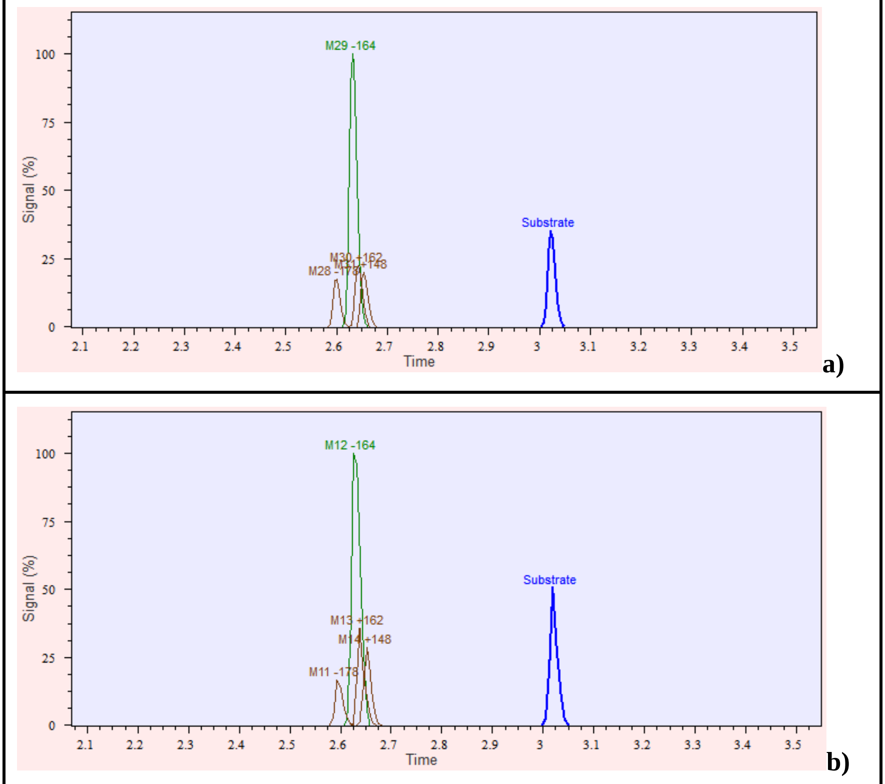


Figure 2. Extracted ion chromatograms for detected verapamil metabolites after 120 minutes of incubation with rat hepatocytes with rat bile: a) MS^E; b) HDMS^E. Blue peak - parent compound; Green peaks - first generation metabolites; Brown peaks- second generation or higher;



The second step of the algorithm assigns chemical structures to the identified metabolites. The software predicts the theoretical fragment ions for the parent compound and metabolites and compares them with the experimentally generated fragment ions. The metabolite fragment ions can have the same *m/z* as a parent fragment ion (the ion is conserved) or can have a defined mass shift (the ion is shifted), respectively. The assigned structures of the previously found metabolites for verapamil are presented in Figures 3 and 4.

Figure 3. Proposed metabolites of GLP-1 found in 120 min incubations.

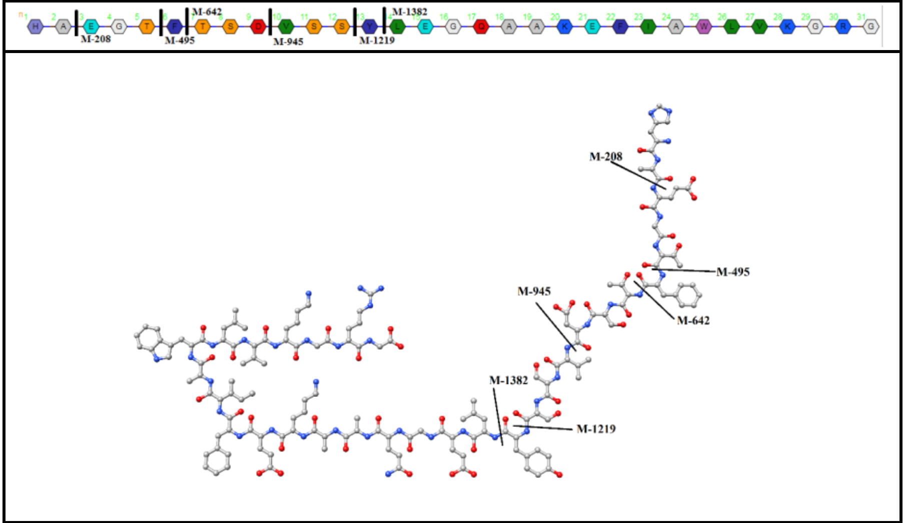
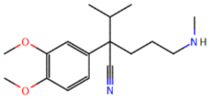
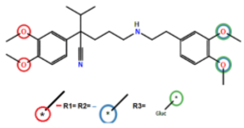
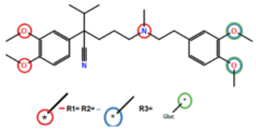
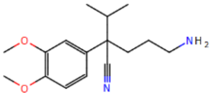
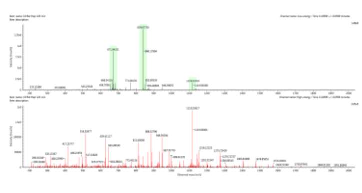
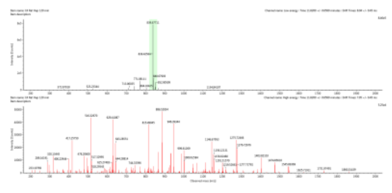
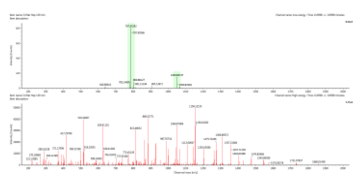
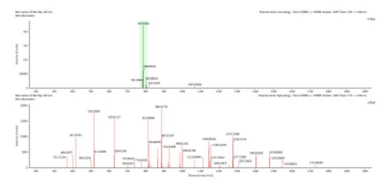

H A E G T F T S D V S S Y L E G Q A A K E F I A W L V K G R G
M-208
M-495
M-642
M-945
M-1382
M-1219

Figure 4. Proposed metabolites of verapamil found in 120 min incubations.

M-164	**M+148**	**M+162**
		
	M-178	
		

MS^E spectra were acquired with a collision energy ramp and no quadrupole selection, the association of fragment and precursor ions is achieved by matching retention time and peak shape. As with MS^E, HDMS^E has no formal preselection of the ions, but precursor and product ions can be correlated based on a shared drift time in the ion mobility cell, resulting in cleaner MS and MS^2 spectra. This is evident upon comparison of the MS^E and HDMS^E spectra generated by UNIFI for GLP-1 and the M-208 metabolite shown in Figure 5. Here HDMS^E has drift separated the multiply charged ions resulting in HDMS^E MS^2 spectra for the drift window containing m/z 839.4253 and m/z 787.4014 for GLP-1 and M-208, respectively.

Figure 5. MS^E and HDMS^E spectra for GLP-1 and metabolite M-208.

	MS^E	**HDMS^E**
GLP-1		
M-208		

The MS/MS spectra obtained using MS^E and HDMS^E for the GLP-1 and the metabolite M-495 peak which elutes at 2.17 minutes and the metabolite M-642 peak which elutes at 2.15 are shown in Figure 6 and 7, respectively. Red ions denote substrate fragments that match with metabolite fragments. Blue ions denote substrate fragments that do not match with metabolite fragments or that the fragmentation process does not follow any of the rules used in fragment generation. There are 47 matching fragments in the MS^E spectrum, with 17 mismatches and 10 matching fragments in the HDMS^E spectrum with only 3 mismatching fragments identified for metabolite M-495. For the metabolite M-642 there are 51 matching fragments and 12 mismatches were identified in the MS^E and 21 matching fragments and no mismatches in the HDMS^E. For example, in MS^E m/z 946.39 and m/z 643.29 were found in both MS^2 spectra for metabolite and substrate but they did not match the proposed metabolite structure and so they were colored in blue. In HDMS^E m/z 946.39 and m/z 643.29 were found only in the MS^2 of substrate and were not found for the metabolite, supporting the proposed metabolite by removing the mismatch. The difference between the observed and theoretical (exact) *m/z* was less than 3 ppm for both metabolites. The MS/MS spectra obtained using MS^E and HDMS^E for the GLP-1 and other selected metabolites were provided in Supplementary materials (SFig. 3-6). For all selected metabolites, HDMS^E was cleaner with less ions in total and along with the number of mismatches. The less observed ions and cleaner spectra was a result of drift separation of different charge states from the same molecule and drift separation of chromatographically unresolved peaks.

For each metabolite a score is calculated and reported. This score is based on the number and intensity of matches/mismatches between the theoretical fragment *m/z* value and the *m/z* value observed in the fragment ion spectrum. A high score denotes an assignment which is more likely to be correct due to a high number of matching fragments, a low number of mismatching fragments, and a low average *m/z* difference between the observed and computed values (<3 ppm). However, the ratio of structurally matched to mismatched product ions should be considered since its increases confidence in the proposed metabolites structures. This metabolite has been previously reported when we performed metabolite identification for GLP-1 incubated with dipeptidyl peptidase-4 and neprilysin and used DDA methodology, with full scan/data-dependent MS/MS analyses on a Q Exactive™ Hybrid Quadrupole-Orbitrap Mass Spectrometer (Thermo Fisher Scientific). We obtained similar metabolite structure, kinetic analysis results and similar results for the matches/mismatches score with 52 matches and 2 mismatches [12]. For the DDA data Mass-MetaSite combines MSMS spectra for all charge states in one. While processing MS^E and HDMS^E data Mass-MetaSite selects the spectra with a drift time of the most intense peak of *m/z* of one of the charges. The combination of different charge states cannot be done at this moment. Thus, we cannot confirm if the larger number of matches is caused by combining MS^2 spectra from multiple charge states, or that the

DDA experimentation generated a larger number of ions, or a combination of these two reasons.

Figure 6. MS^2 spectra for the GLP-1 metabolite M-495.
Red peaks - correlated with fragments that match between metabolite and parent; Blue peaks - correlated only with parent;

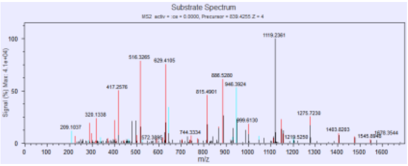
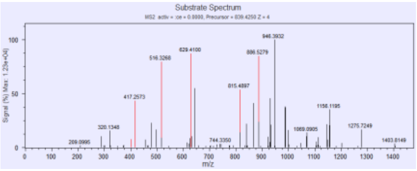
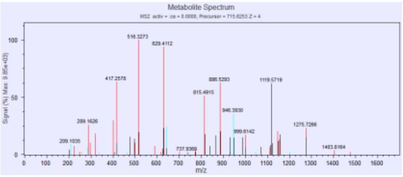
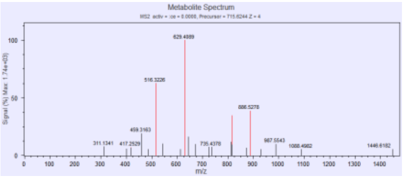
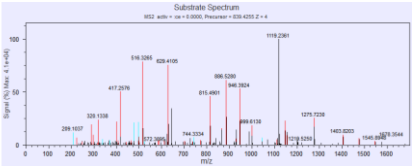
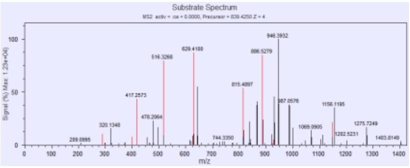
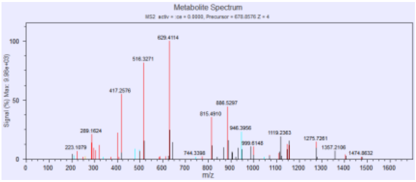
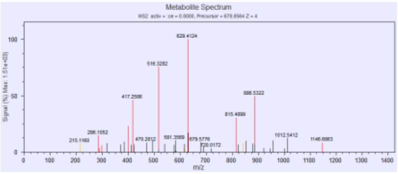
MS^E	**HDMS^E**
	
	

Figure 7. MS^2 spectra for the GLP-1 metabolite M-642.
Red peaks - correlated with fragments that match between metabolite and parent; Blue peaks - correlated only with parent; Orange peaks - correlated only with metabolite;

<table>
<tr><th>MS^E</th><th>HDMS^E</th></tr>
<tr><td>


</td><td>


</td></tr>
<tr><td>


</td><td>


</td></tr>
</table>

In Tables 4 and 5 the number of matching and mismatching fragment ions is reported for selected metabolites of GLP-1 and verapamil, respectively for MS^E and HDMS^E data. The total number of resolved fragment ions for GLP-1 was lower for HDMS^E corresponding to the enhanced precursor selectivity obtained with drift time-resolved product ion spectra. While the absolute score is lower for the metabolites elucidated using HDMS^E data for GLP-1 and verapamil, which is a function of the lower total amount of matched and mismatched fragments, the structural assignment is expected to be more reliable, because the number of mismatched ions is significantly lower for HDMS^E data. A mismatch represents the existence of a fragment ion in the spectra that is not rationalized by the proposed structure. Therefore, although the score for MS^E is higher for the GLP-1 metabolites, the number of mismatches for HDMS^E was lower for all the metabolites, thus the proposed metabolites structures are more reliable.

Table 4. Number of match/mismatches, and structural assignment score based on MS^E and HDMS^E spectra for the GLP-1 metabolites

	MS^E				HDMS^E			
	Match	**Mis-match**	**Match-Mis-match Ratio**	**Score**	**Match**	**Mis-match**	**Match -Mis-match Ratio**	**Score**
M-1219	33	2	16.5	1705	21	0	21*	1082
M-1382	13	0	13*	851	10	0	10*	241
M-945	20	1	20	1265	13	1	13	942
M-495	47	17	2.8	1715	10	3	3.3	819
M-642	51	12	4.2	1808	21	0	21*	1072

Table 5. Number of match/mismatches, and structural assignment scores based on MS^E and HDMS^E spectra for the verapamil metabolites

	MS^E				HDMS^E			
	Match	**Mis-match**	**Match-Mis-match Ratio**	**Score**	**Match**	**Mis-match**	**Match-Mis-match Ratio**	**Score**
M-164	8	2	4	399	10	1	10	557
M+162	11	3	3.7	539	9	1	9	876
M+148	13	3	4.3	513	10	3	3.3	601
M-178	5	0	5*	395	4	2	2	22

A metabolic stability or clearance experiment has multiple time points. Therefore, a software system should be able to compare the multiple samples of an experiment and perform kinetic analyses. Metabolites from each sample were grouped together with the metabolites from the other samples by “analysis clustering” performed in WebMetabase. The appearance of the metabolites and the disappearance of the parent represented as the areas of the found peaks against the incubation time were plotted and compared. The substrate and metabolite time profiles for the major metabolites of MS^E and HDMS^E data of GLP-1 and verapamil are shown in Supporting Figures 1 and 2, respectively. The kinetic analyses in HDMS^E and MS^E data was very similar, even when the concentration or signal of the metabolite was very low. Therefore, both MS^E and HDMS^E provide similar kinetic results, with similar sensitivity.

Conclusion:

To our knowledge this is the first time that an HDMS^E approach has been published in the field of metabolite identification of peptides. Here we found that HDMS^E outcomes were comparable to the previously published DDA outcomes with similar values between the numbers of structural matching and mismatching fragment ions. A comparison of DIA approaches (MS^E and HDMS^E) resulted in the same metabolite structures for both GLP-1 and verapamil. MS^E data acquisition generated more matching and mismatching ions due to unresolved charge states and chromatographic peaks. The ratio of structurally matched to mismatched product ions found by Mass-MetaSite was greater HDMS^E improving confidence in the structures proposed through the addition of ion mobility-based data acquisitions. Moreover, HDMS^E displayed the ability to drift separate charge states and unresolved chromatographic peaks, evident by the lower number of mismatching fragments. Future work will be aimed toward combining the HDMS^E acquired MS^2 spectra of peptides from the multiple drift windows associated with the multiple charge states of peptides to improve structure identification.

Acknowledgements:

We are grateful to Alessio Zammataro, Jeff Goshawk and Martin Lunt for their support with integration of UNIFI and Mass-MetaSite/WebMetabase. We are also grateful to Jared Borher and

Rich Gundersdorf for helping to generate samples and to Dr. Kevin Bateman for continual discussions on testing and implementation of new technologies.

References:

1. Dwivedia P, Schultzb A. J, Hill H. H. Jr Metabolic profiling of human blood by high-resolution ion mobility mass spectrometry (IM-MS). International Journal of Mass Spectrometry. 298 (2010) 78–90.
2. Katsila T, Siskos AP, Tamvakopoulos C. Peptide and protein drugs: the study of their metabolism and catabolism by mass spectrometry. Mass Spectrom Rev. 2012;31(1):110-133. doi: 10.1002/mas.20340.
3. Cui Q, Lewis IA, Hegeman AD, Anderson ME, Li J, Schulte CF, Westler WM, Eghbalnia HR, Sussman MR, Markley JL: Metabolite identification via the Madison Metabolomics Consortium Database. Nat Biotechnol 2008, 26(2):162–164.
4. Last RL, Jones AD, Shachar-Hill Y: Towards the plant metabolome and beyond. Nat RevMol Cell Biol 2007, 8:167–174.
5. Patti GJ, Yanes O, Siuzdak G: Innovation: Metabolomics: The apogee of the omics trilogy. Nat RevMol Cell Biol 2012, 13(4):263–269.
6. Bateman, K.P. et al. (2007) MSE with mass defect filtering for in vitro and in vivo metabolite identification. Rapid Commun. Mass Spectrom. 21, 1485–1496
7. Goodwin CR, Fenn LS, Derewacz DK, Bachmann BO, McLean JA. Structural Mass Spectrometry: Rapid Methods for Separation and Analysis of Peptide Natural Products. J Nat Prod. 2012 Jan 27;75(1):48-53. doi: 10.1021/np200457r
8. Distler U, Kuharev J, Navarro P, Levin Y, Schild H, Tenzer S.:Drift time-specific collision energies enable deep-coverage data-independent acquisition proteomics. Nature methods. V.11 No.2.Feb 2014. 167-175. doi:10.1038/nmeth.2767

9. Lapthorn C, Pullen F, Chowdhry BZ. Ion mobility spectrometry - mass spectrometry (IMS-MS) of small molecules: separating and assigning structures to ions.Mass Spectrom Rev. 2013 Jan-Feb;32(1):43-71. doi: 10.1002/mas.21349.
10. Bonn B, Leandersson C, Fontaine F, Zamora I. Enhanced metabolite identification with MS(E) and a semi-automated software for structural elucidation. Rapid. Commun. Mass. Spectrom. 2010;24(21):3127-3138. doi: 10.1002/rcm.4753.
11. Brink A. Metabolites: structure determination and prediction Software aided approaches to structure-based metabolite identification in drug discovery and development. Drug Discov. Today Technol. 2013:10(1): e207-217. doi: 10.1016/j.ddtec.2012.12.001.
12. Radchenko T, Brink A, Siegrist Y, Kochansky C, Bateman A, Fontaine F, et al. (2017) Software-aided approach to investigate peptide structure and metabolic susceptibility of amide bonds in peptide drugs based on high resolution mass spectrometry. PLoS ONE 12(11): e0186461.
13. Sharma R, McDonald T. S, Eng H. In Vitro Metabolism of the Glucagon-Like Peptide-1 (GLP-1)–Derived Metabolites GLP-1(9-36) amide and GLP-1(28-36) amide in Mouse and Human Hepatocytes. Drug Metab Dispos 41:2148–2157, December 2013.
14. Reder-Hilz B,·Ullrich M, Ringel M, Hewitt N, Utesch D, Oesch F, Hengstler J. G. Metabolism of propafenone and verapamil by cryopreserved human, rat, mouse and dog hepatocytes: comparison with metabolism in vivo. Naunyn-Schmiedeberg's Arch Pharmacol (2004) 369: 408–417
15. Iwamoto N, Shimada T. Recent advances in mass spectrometry-based approaches for proteomics and biologics: Great contribution for developing therapeutic antibodies. Pharmacology and Therapeutics 185 (2018) 147–154.
16. http://www.waters.com/waters/en_US/MetaboLynxXS/nav.htm ?cid=513803&locale=en_US
17. https://sciex.com/products/software/metabolitepilot-software

18. Sun L, Zhang S Q, Zhong D F. In vitro identification of metabolites of verapamil in rat liver microsomes. Acta Pharmacol Sin. 2004 Jan;25(1):121-8.
19. Bateman K.P, Kellmann M, Muenster H, Pappa R, Taylor L. Quantitative–Qualitative Data Acquisition Using a Benchtop Orbitrap Mass Spectrometer. J Am Soc Mass Spectrom 2009, 20, 1441–1450. doi:10.1016/j.jasms.2009.03.002
20. Gallagher R, Dillon L, Grimsley A, Murphy J, Samuelsson K, Douce D. The application of a new microfluidic device for the simultaneous identification and quantitation of midazolam metabolites obtained from a single micro-litre of chimeric mice blood. Rapid Commun Mass Spectrom. 2014;28(11):1293-1302. doi:10.1002/rcm.6902.
21. Walles M, Thum T, Levsen K, Borlak J. Metabolism of verapamil: 24 new phase I and phase II metabolites identified in cell cultures of rat hepatocytes by liquid chromatography-tandem mass spectrometry. J Chromatogr B Analyt Technol Biomed Life Sci. 2003 Dec 25;798(2):265-74. https://doi.org/10.1016/j.jchromb.2003.09.058 .

Supporting Table 1. Mass-MetaSite settings are shown with experimental details for GLP-1

Mass-MetaSite Settings		
Import	Protonation policy	pH=7
	Maximum number of conformers	20
Metabolite generation	Minimum mass	50
	Metabolite stereochemistry and redundant metabolites	ignored
	MIM (the percentage of the monoisotopic mass of the parent)	30%
	Common cytochrome P450 reaction mechanisms	none
Mass settings, experiment	Amide Hydrolysis	true
	Retention time range (min)	not used
	Standard mode	deactivated
	Peptide mode	activated
	GSH mode	deactivated
Mass settings, MS peaks	Maximum metabolite count limit	20
	Peak area threshold (%)	0.50%
	Peak area threshold (absolute)	0
	Peak detection smoothing	Medium
	Isotope pattern filtering tolerance	20
Expected metabolites	Rescue computed DRM peaks	not used
	Split computed DRM peaks	not used
	Adducts	not used
	Dimeric Ions	not used
	Neutral loses	not used
	Multi-Charge Ions	used
	Multi-Charge Ions max z	5
	Unexpected metabolites	included
Mass settings, Met ID	Number of metabolite generations	2
	Compound fragmenting, bond breaking limit	2
	Break metabolites	included
	Break metabolites limit	1
	Even electron	MS and MS/MS
	Odd electron	MS and MS/MS
	N-Oxide	MS
Mass settings, DD-MS/MS algorithms, thresholds	Mass spectrometer	TEC Waters Q-TOF
	Same peak tolerance (amu)	0.01
	Chromatogram automatic filtering threshold	0.95

Supporting Table 1. Mass-MetaSite settings are shown with experimental details for GLP-1

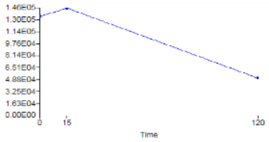
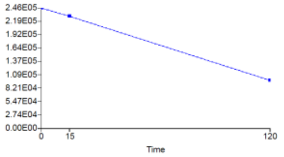
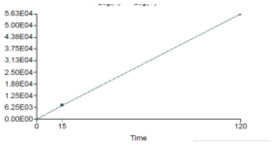
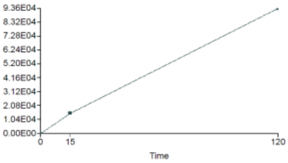
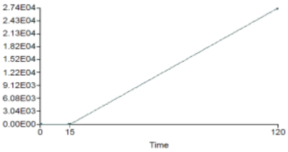
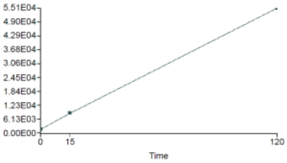
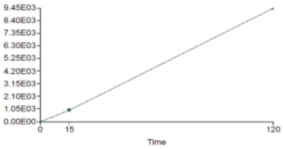
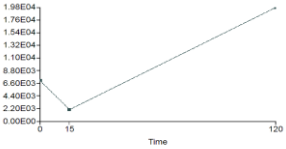
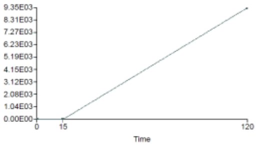
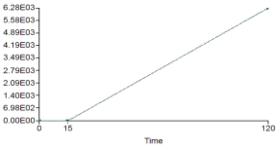
	MS automatic filtering threshold	0.95
	MS/MS automatic filtering threshold	0.95
	Ionization mode	positive $[M+H]^+$
	Signal filtering for MSE data	100
	Signal filtering for HDMSE data	automatic
	Scan filtering	automatic

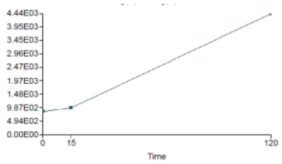
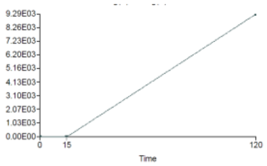
Supporting Table 2. Mass-MetaSite settings are shown with experimental details for Verapamil

Mass-MetaSite Settings		
Import	Protonation policy	pH=7
	Maximum number of conformers	20
Metabolite generation	Minimum mass	50
	Metabolite stereochemistry and redundant metabolites	ignored
	MIM (the percentage of the monoisotopic mass of the parent)	30%
	Common cytochrome P450 reaction mechanisms	none
Mass settings, experiment	Hepatocytes	true
	Retention time range (min)	not used
	Standard mode	activated
	Peptide mode	deactivated
	GSH mode	deactivated
Mass settings, MS peaks	Maximum metabolite count limit	20
	Peak area threshold (%)	0.50%
	Peak area threshold (absolute)	0
	Peak detection smoothing	**Medium**
Expected metabolites	Rescue computed DRM peaks	not used
	Split computed DRM peaks	not used
	Adducts	not used
	Dimeric Ions	not used
	Neutral loses	not used
	Multi-Charge Ions	used
	Multi-Charge Ions max z	**5**
	Unexpected metabolites	**included**
Mass settings, Met ID	Number of metabolite generations	3
	Compound fragmenting, bond breaking limit	4
	Break metabolites	Not included
	Break metabolites limit	0
	Even electron	MS and MS/MS
	Odd electron	MS and MS/MS
	N-Oxide	MS
Mass settings, DD-MS/MS algorithms, thresholds	Mass spectrometer	TEC Waters Q-TOF
	Same peak tolerance (amu)	0.01
	Chromatogram automatic filtering threshold	0.95
	MS automatic filtering threshold	0.95

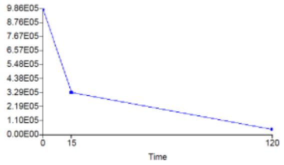
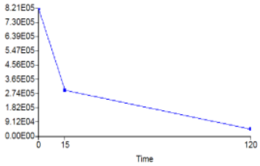
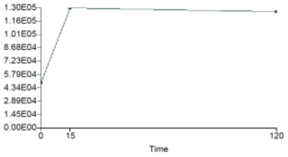
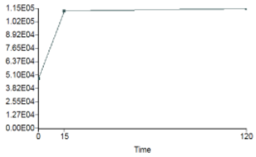
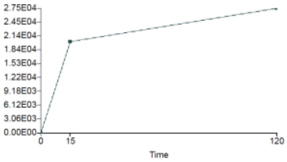
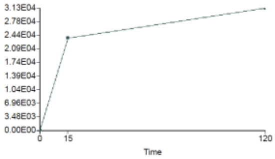
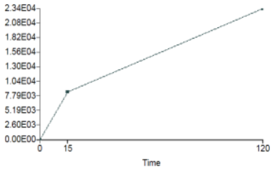
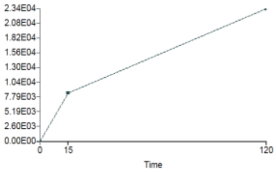
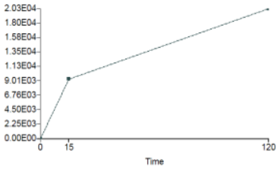
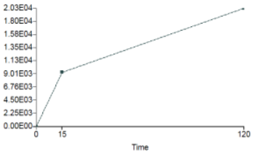
Supporting Table 2. Mass-MetaSite settings are shown with experimental details for Verapamil

	MS/MS automatic filtering threshold	0.95
	Ionization mode	positive $[M+H]^+$
	Signal filtering for MSE data	100
	Signal filtering for HDMSE data	automatic
	Scan filtering	automatic

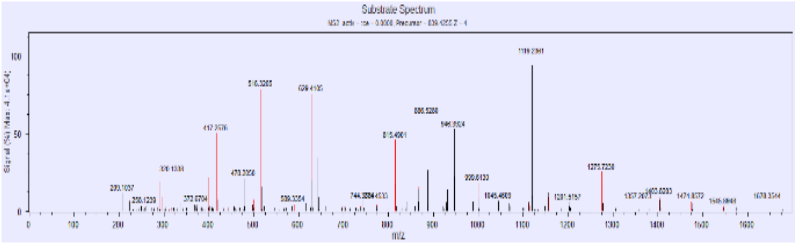
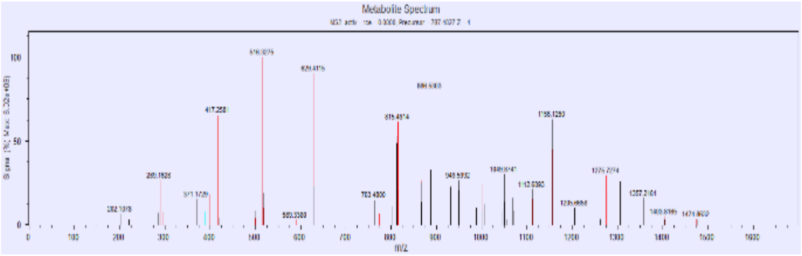
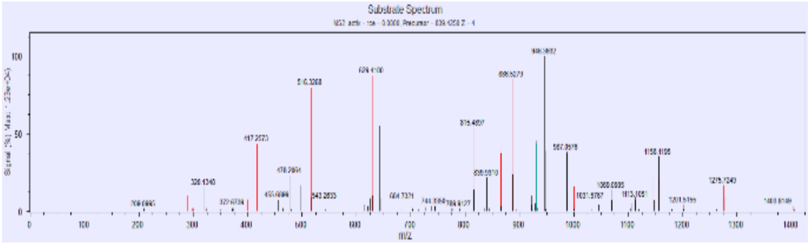
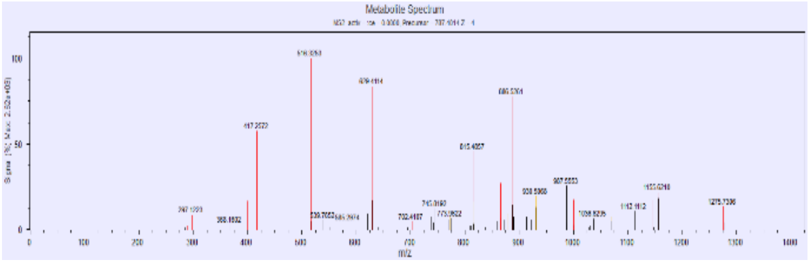
Supporting Figure 1. Peak area as a function of incubation time for GLP-1 and its metabolites.		
	MS^{E}	**HDMS^{E}**
GLP-1		
M-1219		
M-642		
M-945		
M-1382		

<table>
<tr><td colspan="3"><b>Supporting Figure 1. Peak area as a function of incubation time for GLP-1 and its metabolites.</b></td></tr>
<tr><td><b>M-495</b></td><td>
</td><td>
</td></tr>
</table>

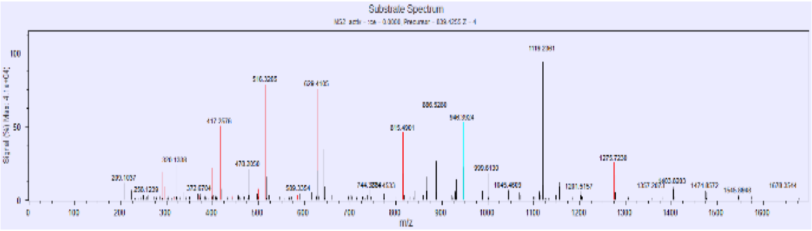
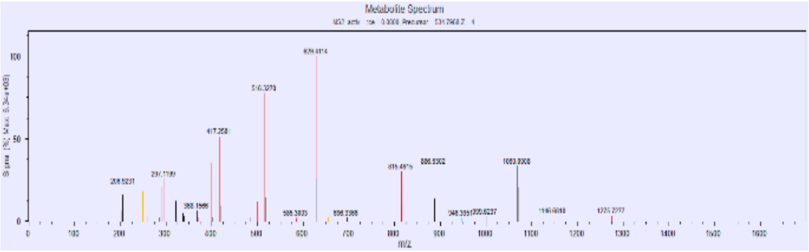
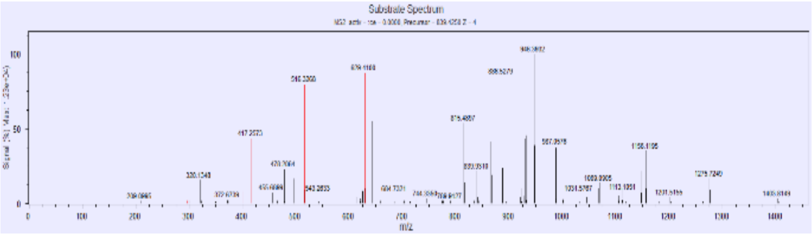
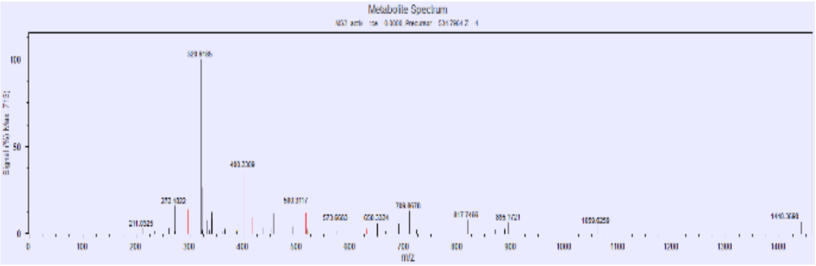
Supporting Figure 2. Peak area as a function of incubation time for verapamil and its metabolites.

	MS^E	**HDMS^E**
Verapamil		
M-164		
M+162		
M+148		
M-178		

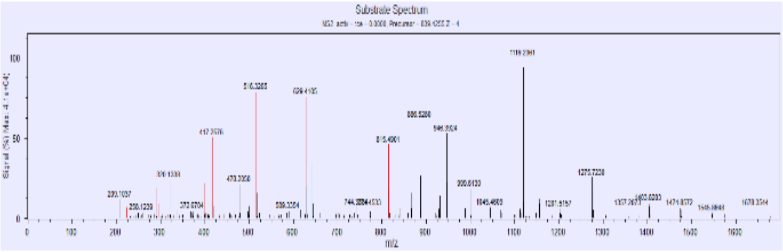
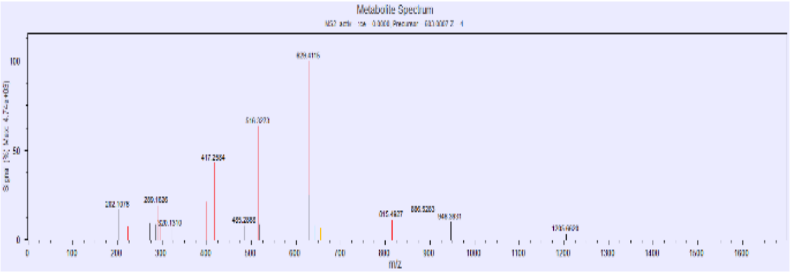
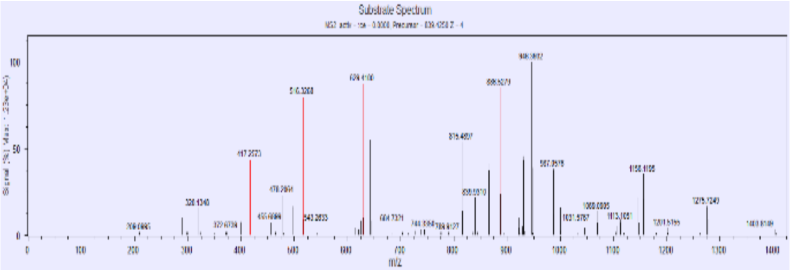
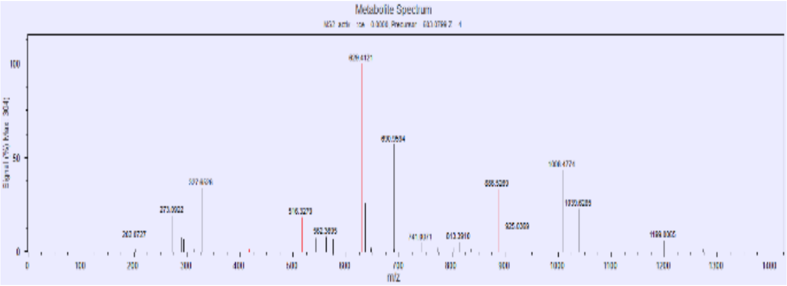
Supporting Figure 3. Full scan/data-dependent MS/MS spectra for the GLP-1 metabolite M-208. Red peaks - correlated with fragments that match between metabolite and fragment; Blue peaks - describe peaks correlated only with parent;

MS^E Substrate	
MS^E Metabolite	
$HDMS^E$ Substrate	
$HDMS^E$ Metabolite	

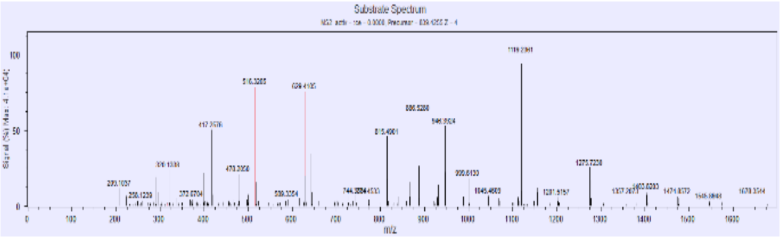
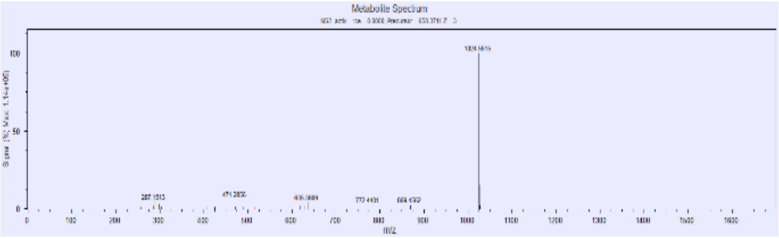
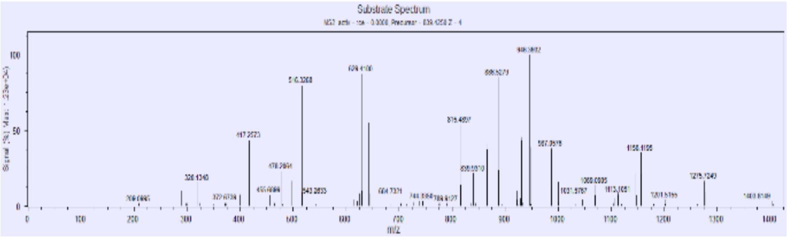
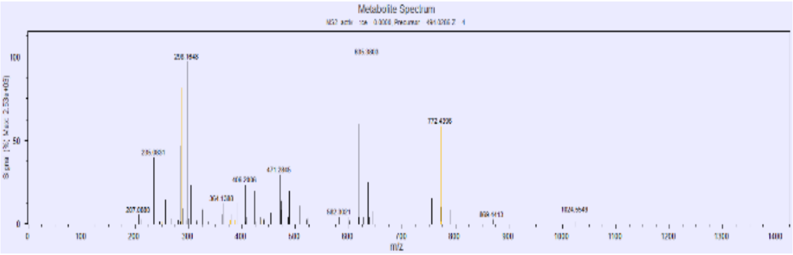
Supporting Figure 4. Full scan/data-dependent MS/MS spectra for the GLP-1 metabolite M-1219. Red peaks - correlated with fragments that match between metabolite and fragment; Blue peaks - describe peaks correlated only with parent;

MS^E Substrate	
MS^E Metabolite	
HDMS^E Substrate	
HDMS^E Metabolite	

Supporting Figure 5. Full scan/data-dependent MS/MS spectra for the GLP-1 metabolite M-945. Red peaks - correlated with fragments that match between metabolite and fragment; Blue peaks - describe peaks correlated only with parent;

MS^E Substrate	
MS^E Metabolite	
HDMS^E Substrate	
HDMS^E Metabolite	

Supporting Figure 6. Full scan/data-dependent MS/MS spectra for the GLP-1 metabolite M-1382. Red peaks - correlated with fragments that match between metabolite and fragment; Blue peaks - describe peaks correlated only with parent;

MS^E Substrate	
MS^E Metabolite	
HDMS^E Substrate	
HDMS^E Metabolite	 Metabolite Spectrum 100 50 0 772.4206 471.2845 0 100 200 300 400 500 600 700 800 900 1000 1100 1200 1300 1400 m/z