

Computational methodologies for DNA-encoded libraries

Dissertation

for the academic degree of

Doctor of Sciences

from the Faculty of Chemistry and Chemical Biology
of Technical University of Dortmund

by

Silvia Chines

Dean: Prof. Stefan Kast

1. Examiner: PD Andreas Brunschweiler
2. Examiner: Prof. Dr. Katja Ickstadt

Nihil ausi, nihil acquiritur.

Nothing ventured nothing gained.

Chi non risica non rosica.

The work presented in this thesis was performed in the time period from July 2018 to June 2022 under the supervision of PD Andreas Brunschweiler at the Faculty of Chemistry and Chemical Biology of the Technical University of Dortmund.

Part of the work presented in this thesis has been published in the following articles:

- **"Screening of metal ions and organocatalysts on solid support-coupled DNA oligonucleotides guides design of DNA-encoded reactions"** M. Potowski, F. Losch, E. Wünnemann, J. K. Dahmen, S. Chines, A. Brunschweiler, *Chem. Sci.* 2019, 10, 10481-10492.
- **"Reaction Development for DNA-Encoded Library Technology: From Evolution to Revolution?"** K. Götte, S. Chines, A. Brunschweiler *Tet. Lett.*, 2020, 61, 151889.
- **"Navigating chemical reactions space - application to DNA-encoded chemistry"** S. Chines, C. Ehart, M. Potowski, F. Biesenkamp, L. Grützbach, S. Brunner, F. van den Broek, S. Bali, K. Ickstadt, A. Brunschweiler (submitted and accepted with revisions).

Table of Contents

1 Kurzfassung.....	5
2 Abstract.....	7
3 General introduction.....	9
3.1 HTS and DNA-encoded libraries.....	11
3.2 DELs technology and challenges.....	13
3.2.1 DEL synthesis.....	13
3.2.1.1 Chemistry.....	14
3.2.1.1.1 DNA-damage.....	15
3.2.1.2 Building blocks.....	16
3.2.2 Selection assay and <i>hit</i> identification.....	16
3.3 Chemoinformatics in drug discovery.....	18
3.3.1 Databases.....	19
3.3.2 Chemoinformatics representation for chemical entities.....	19
3.3.3 Descriptors for similarity and clustering.....	21
3.3.4 KNIME.....	23
4 Aim of the thesis.....	27
5 Chemistry selection.....	29
5.1 Introduction.....	31
5.2 Aim.....	33
5.3 Methods.....	34
5.3.1 KNIME workflow.....	34
5.3.2 Input database.....	34
5.3.3 Data preparation.....	34
5.3.4 Filtering by conditions.....	36
5.3.5 Further refinement.....	36
5.3.6 Reference reactions.....	37
5.3.7 Analysis of the elements.....	37
5.3.8 Reaction descriptors calculation.....	38

5.3.9	Feature extraction.....	38
5.3.10	Reaction classification.....	41
5.3.11	Clustering.....	42
5.3.12	Reactants and catalysts versatility.....	42
5.3.13	Scaffolds in drugs.....	43
5.3.14	Anti-reactions.....	43
5.4	Results and discussion.....	44
5.4.1	Additional findings.....	49
5.4.1.1	Versatile reactants.....	49
5.4.1.2	Metal catalysts analysis.....	51
5.4.1.3	Accessible scaffolds.....	54
5.4.2	Experimental validation of applicable reactions.....	56
5.4.2.1	Pyrrrole synthesis starting with aldehyde, amine and ethylacetoacetate.....	56
5.4.2.2	Pyrrrolidine synthesis starting with aldehyde, aniline and thioester.....	59
5.5	Conclusions.....	61
6	Building blocks selection.....	64
6.1	Introduction.....	66
6.1.1	Chemical space coverage.....	66
6.1.2	Chemical space and molecular properties.....	66
6.1.3	Building blocks selection by chemoinformatic tools.....	69
6.1.4	Chemical space visualization and similarity.....	70
6.1.4.1	Principal component analysis (PCA).....	70
6.1.4.2	Principal moment of inertia (PMI).....	72
6.2	Aim.....	74
6.3	Methods.....	75
6.4	Results and discussion.....	80
6.4.1	Libraries based on Povarov, Biginelli and <i>aza</i> -Diels-Alder reactions.....	80
6.4.2	Library based on reductive amination and Suzuki coupling.....	84
6.5	Conclusions.....	93
7	Hits validation by molecular docking.....	95

7.1	Introduction.....	97
7.1.1	Molecular docking for <i>hit</i> identification.....	97
7.1.2	Critical steps in a docking procedure.....	100
7.1.2.1	Ligands preparation.....	100
7.1.2.2	Protein preparation.....	100
7.1.2.3	Definition of the binding site.....	100
7.1.3	DEL and docking.....	101
7.1.4	Protein selection in DEL.....	101
7.2	Aim.....	108
7.3	Methods.....	109
7.4	Results and discussion.....	111
7.4.1	KNIME workflow.....	111
7.4.1.1	Frequent hitters identification.....	112
7.4.1.2	Molecule generation and labelling.....	114
7.4.1.3	Correlations between building blocks and proteins.....	118
7.4.2	Docking experiments.....	120
7.4.2.1	BCL-X _L	120
7.4.2.2	MKK7.....	128
7.4.2.3	MDM2.....	131
7.5	Conclusions.....	135
8	General conclusions and future perspectives.....	137
9	<i>Experimental part</i>	143
9.1	Chemistry selection.....	143
9.1.1	KNIME workflow and tables.....	143
9.1.2	Procedures for the selected reactions and analytical data.....	143
9.1.2.1	Materials and instruments.....	143
9.1.2.2	General procedures (GP).....	144
9.1.2.2.1	Amide coupling (GP1).....	144
9.1.2.2.2	Pyrrole synthesis on DNA-aldehyde conjugate (GP2).....	145
9.1.2.2.3	Pyrrolidine synthesis on DNA-amine conjugate (GP3).....	146

9.1.2.3 Analytical data.....	146
9.2 Building blocks selection.....	158
9.3 Hits validation by molecular docking.....	162
9.3.1 KNIME workflow.....	162
9.3.2 Molecular Docking.....	167
9.3.2.1 Docking with SeeSAR.....	167
9.3.2.2 Docking with Glide.....	167
10 Abbreviations.....	169
11 Acknowledgment.....	173
12 References.....	175
13 Appendix.....	185

1 Kurzfassung

Die Entdeckung von Arzneimitteln ist ein langwieriger und sehr kosten- und ressourcenintensiver Prozess. Daher wurden Screening-Technologien (physisch und virtuell) eingesetzt, um solche Prozesse zu optimieren und zu beschleunigen. DNA-kodierte Bibliotheken (DELs) haben sich in den letzten 30 Jahren als Alternative zum Hochdurchsatz-Screening herauskristallisiert, da sie zahlreiche Vorteile bieten, wie z. B. Lagerung, Aufreinigung und vereinfachte Handhabung. In einer DEL ist jedes Mitglied mit einer spezifischen DNA-Sequenz markiert, die während des gesamten Verfahrens als Barcode verwendet wird. Diese Eigenschaft ermöglicht die Identifizierung der Moleküle (Hits), welche nach kombinatorischer Synthese und Affinitätstests zusammengeführt wurden. Die DEL-Technologie birgt jedoch drei große Herausforderungen: der Erhalt des DNA-Barcodes, die Abdeckung des chemischen Raums und die Identifizierung der Hits. Diese Faktoren werden durch die Größe der DNA-kodierten Bibliotheken und die daraus resultierende Menge der erzeugten Daten noch erschwert. Daher wurden chemoinformatische Ansätze entwickelt, um die Technologie zu unterstützen, ihre Wirksamkeit zu verbessern und sie zu optimieren.

Diese Arbeit konzentriert sich auf solche Chemoinformatischen Werkzeuge und schlägt neuartige Lösungen zur Unterstützung des Designs, der Datenanalyse und der Validierung der DEL-Technologie vor. Dabei wird hauptsächlich die KNIME Analytics Plattform aufgrund ihrer Einfachheit und Zugänglichkeit für Chemiker und Biologen verwendet.

Im ersten Kapitel (Chemistry selection) wird ein Algorithmus vorgestellt, der den Raum der chemischen Reaktionen kartiert und sortiert, um Reaktionen auszuwählen, die potenziell für die DEL-Synthese verwendet werden können. Solche Reaktionen müssen Anforderungen erfüllen, welche mit der Kompatibilität der DEL-Umgebung und der kombinatorischen Synthese zusammenhängen. Sie müssen ein gewisses Maß an Wasser tolerieren, attraktive Gerüste für die Entdeckung von Arzneimitteln erzeugen und Vielfalt einbringen. Starke Säuren, Basen, Oxidationsmittel und mutagene Reaktanten sind zu vermeiden und die Reaktionstemperaturen sind auf einen DNA-kompatiblen Bereich zu beschränken. Diese Einschränkungen sind wichtig, um die Integrität des DNA-Barcodes zu erhalten und eine optimale Leistung der Technologie zu gewährleisten. Darüber hinaus beschreibt der

Algorithmus die Reaktionen anhand spezieller molekularer Deskriptoren, die den Kern der Reaktion, d. h. den Teil der Reaktanten, der in Produkte umgewandelt wird, berücksichtigen.

Auf der Grundlage solcher Deskriptoren wurden die Reaktionen geclustert, um den chemischen Reaktionsraum abzubilden. Reaktionen werden so in ähnliche und unähnliche geordnet, um den menschlichen Aufwand auf ein Minimum zu reduzieren. Aus den Clustern konnten fesselnde Reaktionen auf der Grundlage der Bedeutung des erzeugten Gerüsts ausgewählt und erfolgreich auf DNA-kodierte Substrate angewendet werden. Auch zusätzliche Informationen könnten so extrahiert werden.

Im zweiten Kapitel (Building blocks selection) wurde die KNIME Analytics Platform eingesetzt, um die Bausteine (BBs) für die Bibliothekssynthese auszuwählen, mit dem Ziel, die strukturelle Vielfalt zu erhöhen. Nach Anwendung der entsprechenden Filter wurden virtuelle Bibliotheken entsprechend der ausgewählten BBs und der etablierten DNA-kompatiblen Chemien aufgelistet. Die Reaktionen wurden virtuell mit den KNIME-Erweiterungen durchgeführt, und die Endprodukte der Bibliotheken wurden hinsichtlich ihrer Vielfalt und Ähnlichkeit mit großen Datenbanken arzneimittelähnlicher Verbindungen untersucht.

Im dritten Kapitel (DNA-encoded library validation by molecular docking) wurden die Hit-Moleküle, die durch den Selektionsassay einer 100.000-gliedrigen Bibliothek erzeugt wurden, mit molekularer Docking-Software validiert. Drei Proteine wurden detailliert beschrieben und ihre jeweiligen Hits wurden als Liganden im Docking-Verfahren verwendet. Vor dem Docking wurden die häufigsten *Hits* identifiziert und die Daten mit KNIME aufbereitet. Das Docking wurde mit zwei verschiedenen Programmen durchgeführt, um die Robustheit zu gewährleisten. Die Anwendung des Dockings als Validierungsmethode erwies sich als nützlich für die Bestätigung von Treffern, für die Vorhersage der Bindungspositionen und für die Empfehlung einer eventuellen QSAR-Modifikation.

2 Abstract

Drug discovery is a long and highly intense process in terms of costs and resources. Therefore, screening technologies (tangible and virtual) have been implemented to optimize and accelerate it. DNA-encoded libraries (DELs) emerged in the last 30 years as an alternative to the high-throughput screening, due the numerous advantages. such as storage, purification and procedural ease. In a DEL, each member is tagged by a specific DNA sequence, which is used as a barcode over the whole procedure. This characteristic enables the identification of the molecules (*hits*) after pooling them all together for the combinatorial synthesis and after the affinity assay. However, the DEL technology presents three main challenges, among others: the preservation of the DNA barcode, the chemical space coverage and the *hit* identification. Such considerations are further complicated by the size of DNA-encoded libraries and the consequent load of produced data. Therefore, chemoinformatics approaches have come to light to support the technology, improving its efficacy and optimizing it.

This thesis focuses on such chemoinformatics tools and proposes novel solutions to aid the design, the data analysis and the validation of the DEL technology, mainly using the KNIME Analytics Platform for its simplicity and accessibility to chemists and biologists.

In the first chapter, an algorithm is reported to chart the chemical reactions space and sort it out in order to select reactions that could potentially be applied to DEL synthesis. Such reactions must fulfil requirements linked to the compatibility to the DEL environment and to the combinatorial synthesis. They must tolerate a certain extent of water, produce attractive scaffolds for drug discovery and introduce diversity. Strong acids, bases, oxidants and mutagenic reactants are to be avoided and the reaction temperatures are to be restricted to a DNA-compatible range. This limitations are essential to preserve the integrity of the DNA barcode in order to ensure an optimal performance of the technology. Furthermore, the algorithm described the reactions according to peculiar molecular descriptors which consider the reactions core, i.e. the reactants part that get converted into products.

Based on such descriptors, the reactions were clustered in order to map the chemical reactions space and order similar and dissimilar reactions, to reduce the human effort to a minimum. From the clusters, captivating reactions could be selected based on the significance of the produced scaffold and successfully applied to DNA-encoded substrates. Additional information could be extracted as well.

In the second chapter, the KNIME Analytics Platform was employed to select the building blocks (BBs) for library synthesis with the purpose of increasing structural diversity. After applying the proper filters, virtual libraries were enumerated according to the selected BBs and established DNA-compatible chemistries. The reactions were virtually performed with the KNIME extensions and the final products of the libraries were investigated in terms of diversity and similarity with big databases of drug-like compounds.

In the third chapter, the *hit* molecules produced by the selection assay of a 100,000 membered library were validated with molecular docking software. Three proteins were reported in details and their respective *hits* were employed as ligands in the docking procedure. Prior to docking, frequent hitters were identified and the data prepared using KNIME. The docking was performed using two distinct programs in order to ensure robustness. Applying the docking as validation method revealed useful for confirming *hits*, for predicting the binding poses and for recommending eventual QSAR modification.

3 *General introduction*

Drug discovery is an indispensable yet complex process and its cost can raise to approximately 2.8 billions. Moreover, due to optimization phases, multiple assessments and clinical trials, on average 12 years are required from the design of experiments until a drug enters the market. [1] The most common routine in drug discovery begins with *in vitro* assays, which produce *hits* and their respective activity or affinity results. Such results are further verified via *in vivo* biological assays, which also serve the optimization of the *hits* to *lead* compounds that represent clinical candidates. Finally, after passing the clinical trials, the molecules can be considered drugs.

3.1 HTS and DNA-encoded libraries

Since the 90's, the process of drug discovery has been dominated by screening technologies, especially *high-throughput screening* (HTS), wherein huge collections of compounds are evaluated for their activity against all kinds of targets. HTS carries two advantages if compared to other conventional pharmacological assays. Firstly, a preliminary structural knowledge about the interrogated protein is not necessary, and additionally, the identified *hit* shows directly activity, without further assessment. [2]

However, the sampling of the chemical space by such collections remains unsatisfactory, because they only partially cover all possible feasible chemical compounds, that is the definition of the chemical space. [3], [4]. It has been estimated that approximately 250,000 dollars are required to screen 1 to 20 compounds in HTS campaign, not to mention the infrastructure necessary to store all compounds separately. [5] Therefore, in recent years other technologies have emerged to overcome these drawbacks, such as DNA-encoded libraries (DELs). DELs are based on a similar principle as *phage-display* technologies, in which the object of study is tagged with unique DNA codes, used as barcodes. [6] This principle brings two main advantages. Firstly, the genetic code enables to differentiate large numbers of unique DNA sequences, and additionally the DNA-encoded substructures can be all stored in one single vial, dramatically reducing the cost of necessary infrastructures. Due to the combinatorial nature of DELs, the library size easily grows to 6-digits numbers. In fact, a one million membered DEL can be constructed via three synthetic steps with 100 *building blocks* per step. Building blocks, or *synthons*, are substructures characterized by appropriate functional groups which enable the connectivity via specific chemistries to build up the final encoded molecules. [7] Furthermore, the

DELs technology presents other advantages due to the very hydrophilic DNA tag, especially in the purification, which can constitute a bottleneck in HTS. [8]

Once DELs are ready, they can be panned over multiple targets in the selection assay. A prototype of the hit identification process is illustrated in Figure 1. In contrast to the HTS, no special optimization needs to be performed in the DEL selection assay as long as the proteins native conformation is preserved. Although only the affinity of the molecules to the target proteins is interrogated in DEL screening assays, this feature remains essential for the biological activity and it furnishes a very good starting point for optimization.

A peculiar step that characterizes DEL only, in comparison to HTS, is surely the sequencing procedure, based today on readily available *next generation sequencing* (NGS) methodologies. [9] During the selection procedure, the molecules with low or no affinity are washed away and the remaining barcode are amplified by PCR (polymerase chain reaction). Hence, NGS is performed to read the DNA barcodes conjugated to the binding molecules. With the sequencing results at hand, the *enrichment factors* (EFs) can be calculated. EFs enable to highlight the most frequent binders, which are thus considered *hits*. Several approaches have been investigated to calculate the EFs, [10] yet one of the most straightforward proposals relies on counting the copies of each sequence and normalize them over their abundance in control experiments. [11] Finally, the selected *hits* are synthesized by conventional organic chemistry in proper amounts for the appropriate biochemical assays.

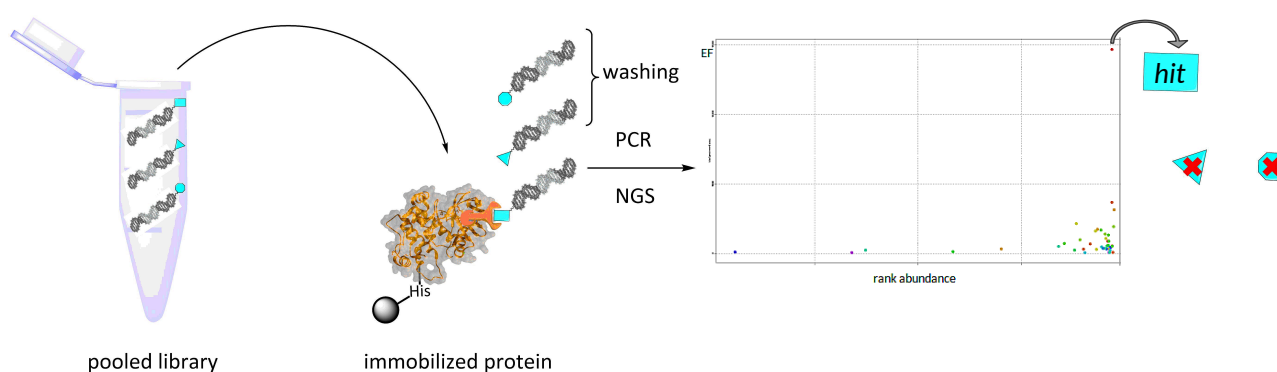


Figure 1: Selection assay with immobilized target and library in solution. The protein is immobilized on the beads via a Histidine tag and the pooled library is incubated with the loaded beads. After washing away the conjugates with low or no affinity towards the target protein, the remaining sequences are amplified and sequenced. The sequencing results are analysed and the hit is identified.

3.2 DELs technology and challenges

3.2.1 DEL synthesis

The most widely used technique for DNA-encoded libraries synthesis is the "*split&pool*" method, which is characteristic of combinatorial libraries. [12] A sequence of DNA that contains the primer for the DNA-polymerase is split in as many vessels as the number of selected building blocks for the first synthetic step. The primer sequence is necessary for the PCR amplification sequential to the selection assay. This general DNA sequence presents on one side a linker with a functional group for initializing the chemical synthesis. On the other side, the primer sequence is ligated with the first set of DNA codes. The chemical reaction involving the first set of building blocks is performed on the linker moiety and the new set of intermediate products are pooled together and split again to ensure that each of them react with each of the following participants of the next step. The alternation of chemical synthesis and DNA sequence elongation is called *cycle*. The second DNA code is ligated to the existing DNA sequence and the same process is repeated for as many cycles as many sets of building blocks are needed until the final library members are completed. The process of synthesizing a three-cycles library is depicted in Figure 2. The DNA barcodes are colored in the same hue as the respective building blocks. Notably, over the synthetic cycles the library size grows exponentially as an effect of its combinatorial nature, as mentioned above. Therefore, chemoinformatics tools to assist the scientist in tracking this process might be of use. In Figure 2, the features of the building blocks (BBS) are highlighted: BBs in the first and second cycle are in general multifunctional as the synthesis must continue, whereas the BBs in the third cycle are monofunctional as they finalize the molecules and reactive groups are to be avoided.

In the context of these synthetic cycles, three essential aspects are to be considered: the ligation procedure, the utilized chemistry and the building blocks. Ligations procedures have been investigated extensively and can be either "*splint*" for single-stranded DNA or "*sticky ends*" for double-stranded DNA. [13],[14] The most challenging decisions are to be made regarding the chemistry and the building blocks. These two factors define to distinct extents the library diversity and, by consequence, the coverage of the chemical space. Notably, libraries based on different chemistries have been demonstrated more effective in *hit* identification than libraries based on one single chemistry but

characterized by diverse *synthons*. [15] For this reason, suitable reactions for DEL synthesis are constantly being developed. [16], [17], [18], [19] However, both factors need to be considered when designing a DEL.

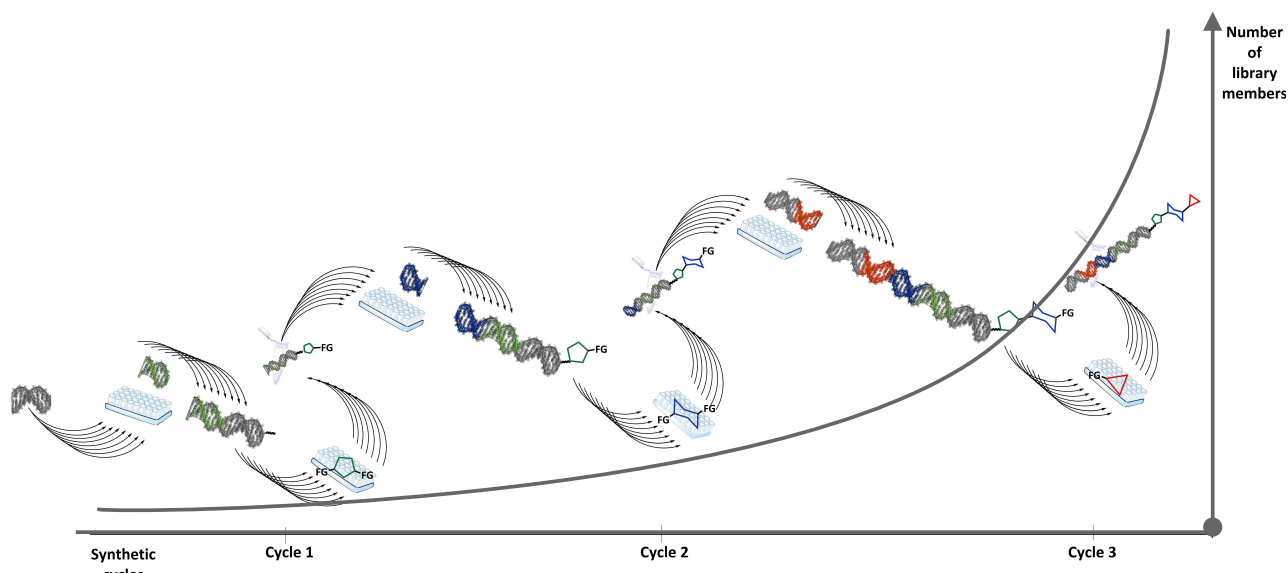


Figure 2: DEL synthesis by split&pool technique. The primer sequence is split in as many vessel as the number of the selected building blocks (BBs) for the first cycle. The first code is ligated to the primer sequence and the coupling with the first set of BBs is performed. All library members are pooled together and split again to repeat the cycle of chemical synthesis and ligation. With the progress in the synthesis the library size grows exponentially due to its combinatorial nature. In the figure, the functionalities of the building blocks are highlighted as well. For the first and the second cycle, multifunctional BBs are required for the coupling with the DNA barcode and for proceeding with the synthesis, respectively. In contrast, the third cycle BBs must be monofunctionalized.

3.2.1.1 Chemistry

DNA-encoded chemistry is a narrow branch of organic chemistry and involve a peculiar kind of reactions, namely robust transformations with a broad substrate scope. [16] These factors are fundamental because in DELs synthesis it is not possible to ensure product formation at each step of the synthesis due to the exponential growth of the library and the pooling steps. In fact, in the analysis of the pooled library, the products can not be distinguished and analysing every single vessel when preparing ultra-large DELs is not efficient. [8] Therefore, robust transformations must deliver product in good yields and must tolerate water, as it constitutes the solvation sphere invariably surrounding the DNA atoms. A broad substrate scope is essential for the chemistry to be applicable in the combinatorial context. Additional criteria could be the connotation as multi-component (MCRs) or heterocycles forming reactions. MCRs implement multiple diversity points in the same step and

heterocyclic scaffolds are widely detected in drugs. [20] Exemplary chemistries that meet those requirements are, for example, the amide coupling or the *click* reaction. [19] However, DNA compatibility represents the most crucial and challenging requirement. This concept is essential in DEL as reactions that harm the DNA or mutate it in any ways would interfere with *hit* identification after sequencing the DNA barcodes at the end of the selection assay, rendering the whole DEL process inefficient.

3.2.1.1.1 DNA-damage

DNA, the acronym for deoxyribonucleic acid, is the physiological macromolecule that encodes all functions of life. The most fascinating feature of DNA is, in fact, the genetic code, formed by four components, the nucleotides. Nucleotides are divided into two pairs according to the respective base, showing pyrimidine (thymidine and cytidine) and purine scaffolds (adenosine and guanosine), respectively. Beside the bases, the nucleotides are composed by the sugar deoxyribose and the phosphate group which connects the sugar substructures. The three-dimensional structure of DNA, the double helix, is arranged by the hydrogen bonds forming between the four bases, precisely adenosine with thymidine and guanosine with cytidine. Those interactions are named Watson and Crick base pairing after the scientists that uncovered them for the first time in 1953. [21] The function of DNA as physiological instructions manual is mediated by the genetic code, therefore its integrity must be preserved. Especially in the field of DNA-encoded libraries, a mutation or any kind of disruption of the DNA used as barcode would be disastrous, as the whole technology is based on reading the DNA barcodes for *hit* identification. Therefore, the chemical reactions employed in the library synthesis must ideally leave the DNA barcode untouched. With this rationale, the effect of various agents on DNA have been explored and some general rules can be summarized as follow. [22] Mutagenic agents are not suitable for DEL as they would irreversibly modify the DNA barcode, impairing the final read-out. Extreme pH values are responsible for *depurination*, on one side, and possible *deamination*, on the other side. In particular, depurination is the acid-mediated mechanism that leads to the loss of the purine bases Adenine and Guanine, while deamination depends on basic conditions and metal-catalysts and provokes mismatching of the DNA bases. [17] Additionally, DNA suffers the presence of strong oxidants which may form the 8-oxo-Guanosine (8-oxo-G) from the Guanosine. A strong oxidant can be considered a chemical entity with oxidation potential higher than 1.29 ± 0.03 V, which is characteristic of the base Guanine at neutral pH. The oxidation has high mutagenic potential because 8-oxoG lesions are more readily depurinated and lead to base-flipping.

Additionally, 8-oxo-G could interact with either the cytosine or the adenosine, transforming the base pair GC into the AT. [23] A more detailed discussion about DNA-encoded chemistries and novel techniques to expand the DEL chemical space is reported in the respective chapter of this thesis.

3.2.1.2 Building blocks

As depicted in Figure 2, building blocks or *synthons* for library synthesis need to show specific properties according to which cycle of synthesis they are employed in. Bi- or tri-functional building blocks are necessary in the first cycle, while mono-functional building blocks are preferred in the eventual third cycle. The factors that influence the selection of building blocks for DEL and some general criteria are further evaluated in the appropriate section of this thesis.

3.2.2 Selection assay and *hit* identification

Once the library is synthesized with satisfying purity, its members can be screened against the selected targets. The choice of the targets to be investigated is arbitrary and depends on the focus of the research group. This subject is discussed more in details in the section *Protein selection in DEL*. The selection assay have been tackled with different approaches over time. At first, soluble proteins were incubated with DELs on solid support, the so called OBOC-DELs (one-bead-one-compound DNA-encoded libraries), and, after the assay, the beads bound to the proteins were identified by a specific antibody. [24] After the advent of the solution phase DEL synthesis by "*split&pool*" technique, the most common approach is to conjugate the target protein on a solid support and incubate it with the DNA-encoded library in solution. In this case, parameters such as the buffer, the target concentration or the beads material play an important role. Testing a range of concentrations and different buffers as well as beads made of varying materials is advisable in order to validate the *hits*. [25] Furthermore, the DEL selection assay has been performed on membrane proteins of living cells or even inside living cells. [26], [27] However, some challenges characterize both techniques. The concentration of the library plays a crucial role for membrane proteins, and the delivery of the library, the crowding effect and the stability of the DNA in intracellular selections assay are to be investigated further.

Independently on the approach, the selection assay bare uncertainties that need validation. A first instance for the purpose of removing false positives is represented by control experiments, performed in parallel to the selection assay. The most common negative control is executed without target, namely by incubating the library with empty beads that do not carry the protein [28], [11] The control

experiments are employed, after the selection assay, in the calculation of the enrichment factors, for identifying hit compounds.

In summary, the uncertainties presented by the DEL selection assay affect the confidence of the read-out after NGS. Therefore, not only statistical methods are necessary to identify *hit* from the sequencing data, but also computational methods are necessary to reduce the noise and distinguish true and false positives as well as true and false negatives. [29]

3.3 Chemoinformatics in drug discovery

As mentioned above, the costs, resources and time required for developing a drug are exorbitant. Therefore, nowadays chemoinformatics tools are utilized to accelerate and optimize this process. [30]

"Chemoinformatics is the mixing of information resources to transform data into information, and information into knowledge, for the intended purpose of making better decisions faster in the arena of drug lead identification and optimisation" [31] This early definition, coined by F. K. Brown in 1998, pointedly describes the purpose and the approach of chemoinformatics in drug discovery. The data produced by chemical or biological experiments are collected and processed to extract information that guide in designing further experiments, optimizing time and resources (Figure 3).

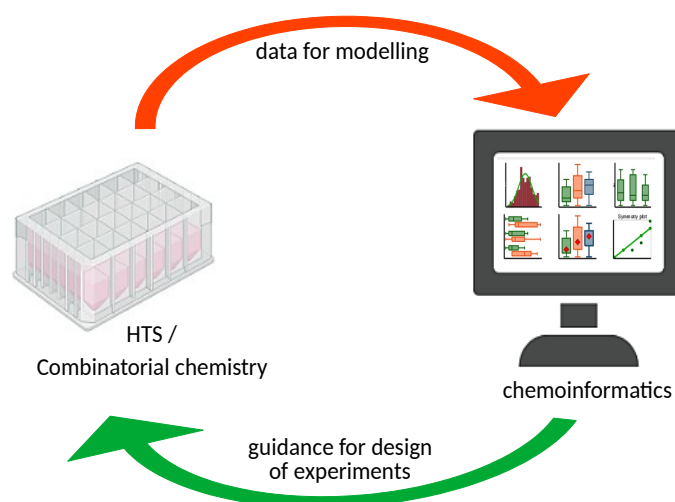


Figure 3: Mutual relationship between experimental procedures and chemoinformatics in the process of drug discovery. Experimental data are employed to predict similar outcomes and improve decision-making in designing experiments.

This purpose is pursued by processing experimental data using statistical methods, in order to predict and better understanding immeasurable chemical phenomena. For example, the three-dimensional structure, confirmed by X-ray crystallography, is known for approximately 1% of the available compounds and, with chemoinformatics methods, it is possible to predict conformations of molecules according to the existing data.[32] The same principle is applicable for predicting chemical properties or activity for molecules which share similarities. [30] Additionally, the exponentially growing amount of data produced by high-throughput techniques, such as HTS or proteomics, can not be handled

without using computers and information technology. [33] These figures highlight the importance of chemoinformatics in an efficient drug discovery process.

3.3.1 Databases

Chemoinformatics is placed at the intersection of information technology and many aspects of chemistry: from synthetic chemistry or analytical chemistry, to physical chemistry or even biochemistry. [34] According to the area of interest, the experimental data can range from chemical structures, to yields, to biological activities, to resolved crystallographic structures of proteins, and such data are employed to generate models for predictions. To facilitate the accessibility for the scientific community, this information is collected and stored into databases. The most commonly known freely available database of molecules with the respective activity data is ChEMBL, which contains 2.2 millions compounds. [35] Other commercial examples are the Enamine screening collection or the Aldrich Market Select with 2.9 and 14 millions molecules, respectively. [36], [37] Such compound collections can be purchased and used in screening campaigns such as HTS or can be used to predict the activity of novel chemical entities. Databases can also be generated virtually and can be used in *virtual screening* (VS), which is the virtual counterpart of HTS. In VS, the binding between target proteins and virtually generated libraries is simulated, in a process called *docking*. A detailed introduction for the docking and its application is reported in the respective section of this thesis. Two main examples of such collections are the GDB-17 (Generated Database) and the ZINC15 with 200 millions and 200 trillions molecules, respectively. [38], [39]

3.3.2 Chemoinformatics representation for chemical entities

In order to predict properties of chemical or biochemical entities, reliable models can be generated based on similar sets. However, the similarity assessment of chemical entities would be a challenging task for a machine if specific machine-readable formats did not exist. With such formats, machines can store and process chemical data, calculate chemical properties and perform similarity search.

SMILES (simplified molecular input line system) is a string format for exemplifying chemical structures. The string format is preferable for processing data in machines because it lacks indents and multi-lines that make reading them more complicated. Atoms are represented by their symbols, bonds by signs according to the type, branches by brackets, and rings by numbers. [40] An exemplary structure with the respective SMILES string is depicted in Figure 4.

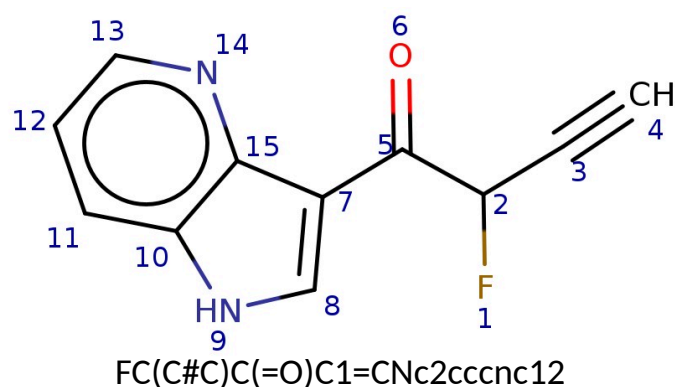


Figure 4: Example of SMILES string: atom symbols appear in uppercase if aliphatic and lowercase if aromatic. Aromatic bonds and single bonds are omitted because considered default, while double bonds and triple bonds are represented by "=" and "#", respectively. Rings are distinguished by Arabic numbers. The blue numbers in the 2D structure represent the numbering followed during the generation of the SMILES string.

Since it uses chemical symbols and clear rules, this format allows for an easy interpretation by machines and by human eyes. However, the sequence of atoms characterizing the string can start with different atoms of the molecules (blue numbers in Figure 4), rendering the conventional SMILES not unique. Therefore, the canonical SMILES have been introduced by Daylight Information System, generating one unique string for each molecule. [41] Sometimes, instead of one exact structure, it is necessary to search for a pattern to detect similarities. For this purpose the format SMARTS (SMILES arbitrary target specification) has been developed by the same company. The same exemplary molecules depicted above is converted in SMARTS format in Figure 5.

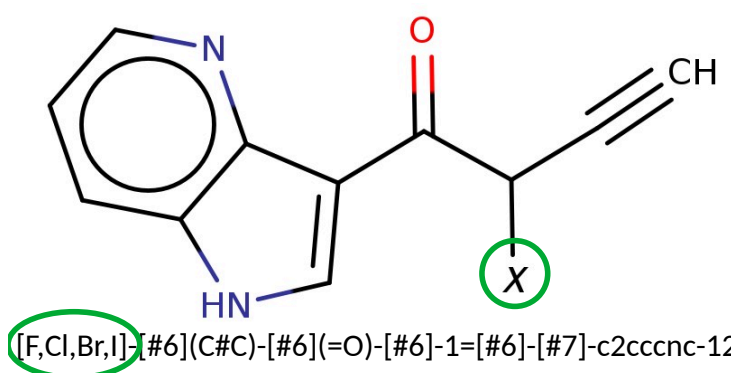


Figure 5: Example of SMARTS pattern: atoms are represented by their atomic number according to the periodic table. The single bonds are explicit, in contrast to the SMILES strings. The X in the 2D depiction represents any halogen and it is translated in SMARTS as a collection of the halogen atoms.

The SMARTS guarantees a very precise specification of atoms and particular features, especially useful in substructure search. [42]

3.3.3 Descriptors for similarity and clustering

Once the machine is able to recognize chemical structures, measuring a sort of distance between molecules is necessary to assess the similarity (or dissimilarity). Such distance can be calculated over descriptors, which are parameters characterizing the molecules. Descriptors can be classified in three categories: one-, two- and three-dimensional. 1D descriptors represent properties of the molecules, such as LogP, number of hydrogen bond donors or acceptors etc., and they account for chemical similarity. 2D and 3D descriptors, on the other hand, account for the structural similarity, because they consider substructures and conformational flexibility, respectively. [34] The issue with structural similarity is that it varies according to the representation, so it is not absolute. Therefore, chemical similarity is often preferred, especially when it is used to compare chemical spaces. A more detailed discussion about 1D descriptors used in drug discovery and their connection with the chemical space is reported in the respective section of this thesis.

In order to observe patterns in the data, draw conclusions and improve predictions for better decisions, grouping similar object is necessary. Clustering is the elected statistical methodology for this purpose. By definition, clustering is an unsupervised classification because it is based on naturally grouping data according to their inherent similarity. [43] Clustering techniques are divided into two main categories: hierarchical and partitional. [44] The hierarchical clustering (HC) can follow two routes: top-down or bottom-up. In the former the whole data set is divided iteratively by dissimilarity, while in the latter the first single object is grouped with other similar objects until the whole dataset is clustered. [45] In Figure 6, the hierarchical clustering dendrogram of the Enamine collection of 646 FDA approved drugs is depicted.

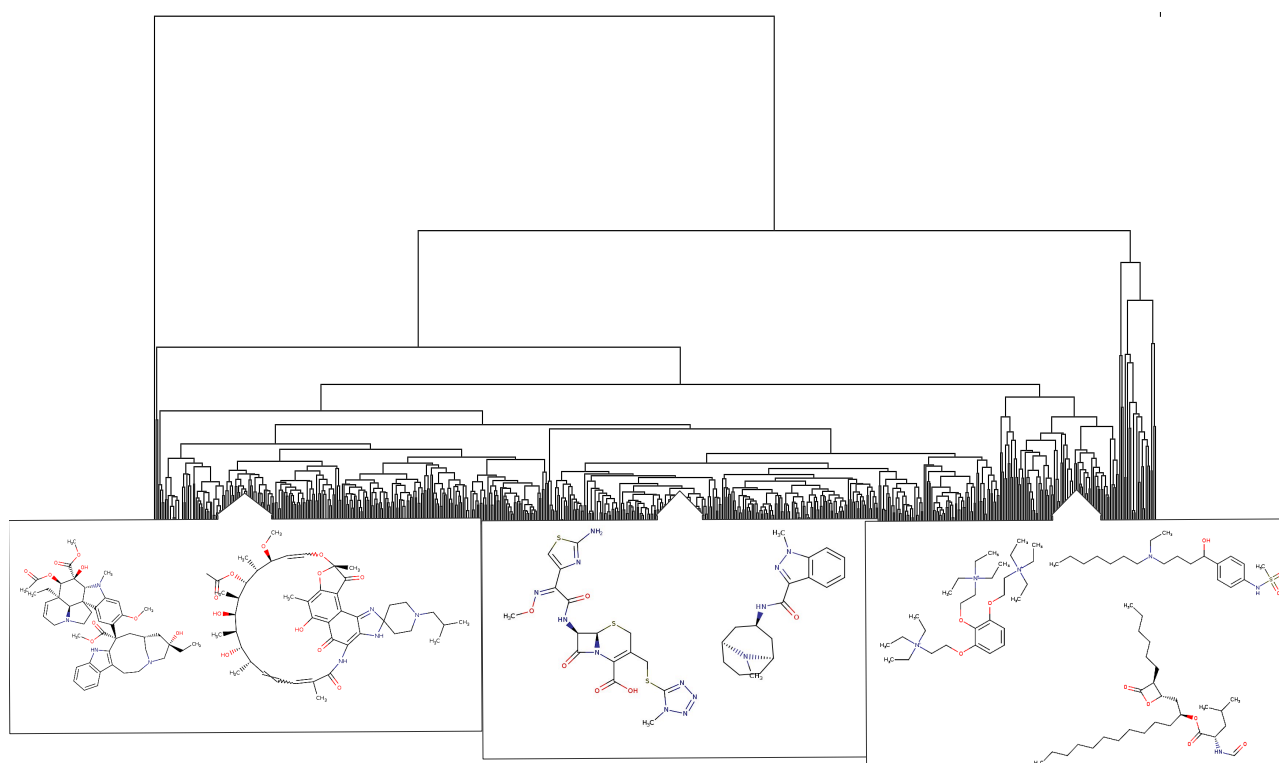


Figure 6: Hierarchical clustering dendrogram and respective affiliated molecules of the Enamine collection of 646 FDA approved drugs. The attributes used for the clustering were the MQN descriptors (please see [section XX](#) of this thesis) and the number of clusters was set to eight. Molecules are clustered according to chemical properties: in the right area of the dendrogram, compounds with long aliphatic chains are grouped, whereas in the left part molecules containing many hydrogen bond donors and acceptors can be spotted. In the centre, molecules with average values of the descriptors are reported.

Due to their complexity, HC approaches require high computational resources and this factor prohibits their usage for large databases. Partitional clustering (PC) are based on the Euclidean distance between objects, assigning close objects to the same cluster. [46] One of the most renowned PC algorithm is the *k-means* which divide the data set in *k* number of clusters. It proceeds by assigning random centroids to each cluster, calculating the distance of each data point to the centroids and grouping them according to the minimum distance. The centroids are then moved and the affiliation optimized until the centroids are stable (Figure 7). [46]

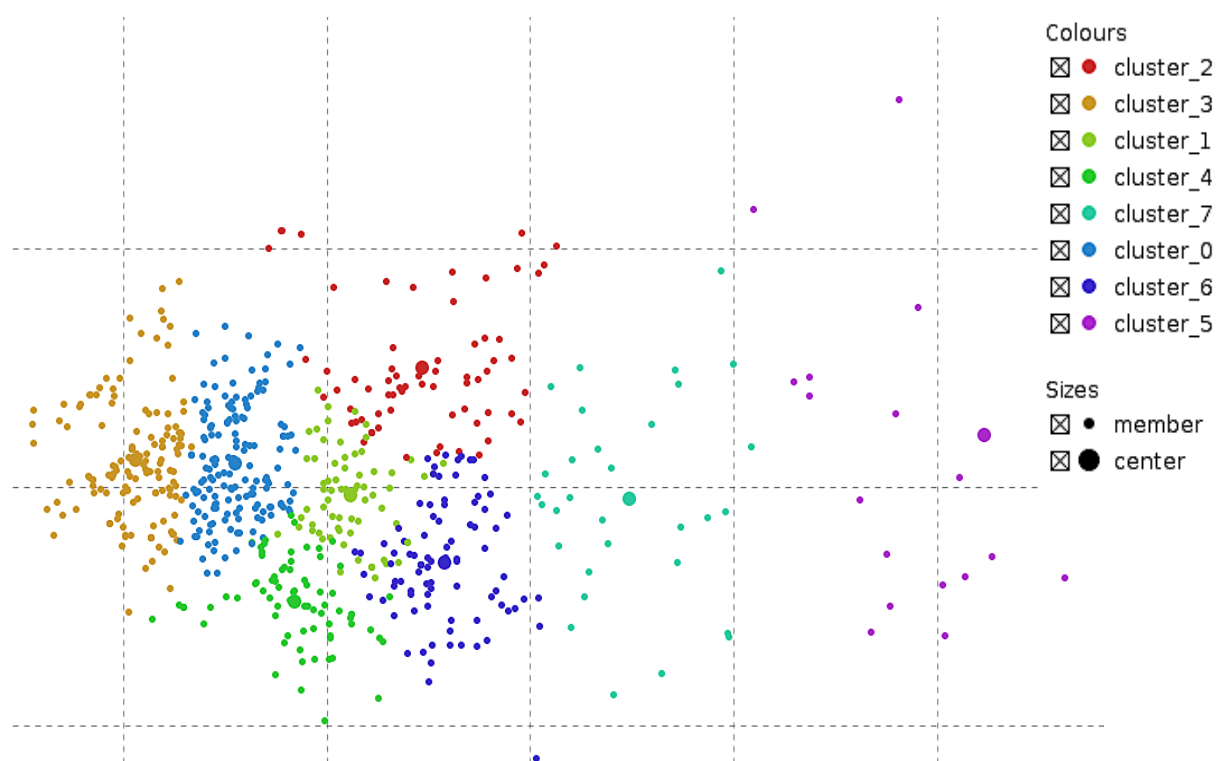


Figure 7: Scatter plot of the k -means clustering on the Enamine collection of 646 FDA approved drugs. The number of clusters, k , was set to eight and the clustering was based on the MQN descriptors. The colors depend on the cluster affiliation and the big data points represent the cluster centres.

However, the k -mean clustering present some limitations: firstly, the number of k and the initial centroids are arbitrary. Additionally, similar to the HC, this method is sensitive to outliers, because values that highly differ from the rest of the data are forcefully assigned to a cluster. For noisy and incomplete data set, the *fuzzy c-means* algorithm is more suitable than the k -means. [47] It is based on the same procedure but the data points can belong to different clusters.

Clustering represents one of the most exploited intersection between chemistry and statistics, as it can be used to group chemical objects, as well as conformations of molecules or chemical reactions. [48]

3.3.4 KNIME

Due to the increasing need for data analysis skills to improve decisions quality, platforms with graphical interfaces, such as KNIME, have been developed to process data and retrieve useful information. [49] KNIME is an open source software for building pipelines (*workflows*) of functions (*nodes*), in a modular fashion: the output of one function becomes the input of the following. This concept can be defined as *graphical programming* and it revealed extremely useful for scientists lacking prior knowledge in conventional programming. The nodes can be divided by functions:

General introduction

input/output nodes read or write data files, processing nodes apply functions to data in order to perform an analysis, visualization nodes allow for graphical reproduction of the results, with graphs or plots. Loop nodes enable iterating the same operation over multiple data according to some parameters. In the KNIME interface (Figure 8) there are several sections, including a Node Repository, to store all available nodes, and the Description section, which introduces each node, to facilitate novices.

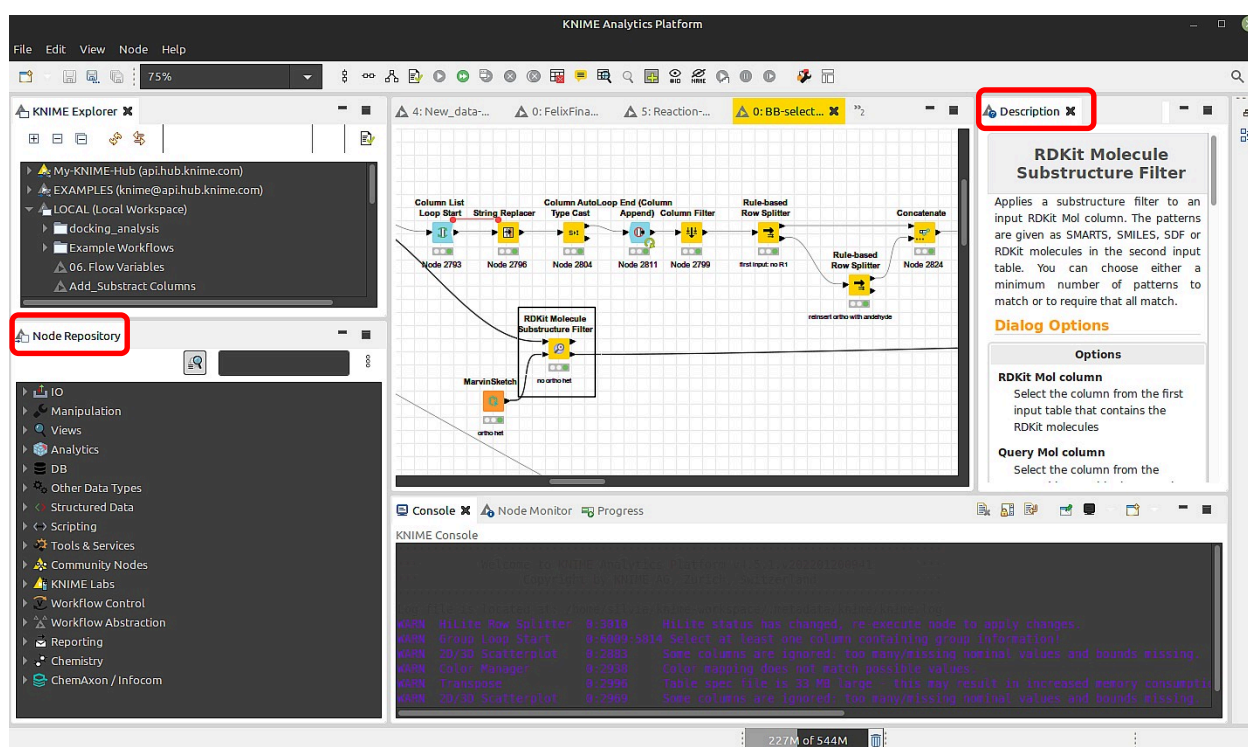


Figure 8: KNIME interface. In the centre, the workflow can be constructed by connecting nodes, which can be searched in the Node Repository, highlighted in red. On the top right, the Description section introduces the selected node.

Moreover, chemistry-oriented extensions can easily be installed, allowing for reading chemical formats, such as SMILES and SMARTS, performing chemical reactions and processing chemical data in general. The most important open source extensions for this purpose are RDKit, Indigo and CDK, [50], [51] whereas the ChemAxon extension is partially available for academic use, especially the Marvin package. [52] Finally, one of the most attractive aspect of KNIME is the constant assistance by the community, guaranteed via the KNIME Hub and Forum. [53]

KNIME and its extensions have been successfully employed within drug discovery projects [54], [55], allowing chemists or biologists for mining their experimental data. KNIME is particularly useful when

dealing with large data sets produced by combinatorial libraries, such as DELs, and with selection assay data. To date, scarce software exists which focus on single aspects of the DEL technology and they are further analysed in the devoted sections of this thesis. However, no unique platform is available to support the DEL technology from design to validation.

|

|

4 Aim of the thesis

Due to its combinatorial nature, the DNA-encoded library technology presents the challenge of dealing with big data sets of either molecules or sequencing data, that can not be handled manually. Analytical platforms and methods can be employed to automatize such process. In this thesis, the use of chemoinformatics methodologies for supporting DEL design is explored. The design of a DNA-encoded library includes the selection of the chemistry and the building blocks accordingly. On the one hand, the chemistry must be DNA-compatible and possibly generate drug-like scaffolds. On the other hand, the building blocks must be adapted to the chemistry and present specific functionalities depending on the synthetic cycle in a given DNA-encoded library synthesis. Furthermore, they should enable covering a large portion of chemical space, to increase the probability of hit identification. Both chemical reactions and building blocks can be selected from chemistry databases according to case-dependent criteria.

Additionally, the validation of DEL hits is currently a pressing topic in medicinal chemistry. [8] Computational techniques such as docking can be employed to reduce the noise in analysing the sequencing data. For example, docking can support in identifying false positives due to unspecific interactions within the selection assay. A careful analysis of the docking results might serve the purposes of improving the selection assay and of predicting the binding poses of the elected hits.

In the ideal case, the whole procedure would be developed via a common platform, namely KNIME. Such platform would allow for computer-assisted DEL design and validation, operated by chemists and biologists without programming skills. This set up would optimize the design of the library, reducing required times and resources to a minimum. In addition, it would increase the effectiveness of the libraries by improving the hit identification rate.

5 Chemistry selection

5.1 Introduction

One of the challenges characterizing DNA-encoded libraries is the coverage of chemical space and it can be tackled by diversifying the library members, by structure and by chemical properties. It has been confirmed computationally that forming different scaffolds using a variety of reactions affects the diversity of the library more than combining different building blocks using the same reaction. [56], [15] In recent years, large efforts have been made in developing DNA-compatible reactions. [8], [16], [17] Beside applying chemical transformations on DNA-tagged substrates and adapting the procedures to conditions that can be tolerated by the oligonucleotides (see the DNA damage section of the general introduction), other approaches include protecting the DNA covalently by a solid support, [22], [57] by electrostatic interactions with resins [58], by micelles [59] or by using PNA [60]. These techniques enable the use of various reaction conditions on DNA-encoded substrates. However, the question remains: which reactions provide DNA-compatible conditions that form relevant scaffolds for screening libraries in a drug discovery project?

In response to the advances in the DEL technology, mapping the chemical reaction space and identify DNA-compatible reactions would greatly aid the design of DELs. Nowadays, the identification of potential reactions to test and apply to DEL design would require sorting hundreds of scientific articles, if algorithms such as the *eDESIGNER* did not exist. [61] This method is comprehensive because it helps in the selection of reactions and building blocks for DELs. However, it considers only established DNA-encoded chemistries, neglecting all the remaining reactions that would allow for venturing outside the beaten track of DEL chemistry.

Several tools have been developed to predict chemical reactions, especially to optimize the retrosynthetic route knowing the target molecule. Two noteworthy examples are *Chematica*, developed in the Grzybowski's group [62] and the work of Segler *et al.* [63]. These computational tools are able to identify the best synthetic procedures in terms of speed and resources, surpassing the human counterpart. Further essential aspects of reactions prediction are optimizing the conditions and eventually balancing reactions in a database, addressed by Gong *et al.* with *DeepReact+* [64] and by Delannée *et al.* with the *ReactionCode* [65], respectively. The former predicts yield and optimal conditions for the input reactions, whereas the latter unifies the reactions format, enabling the classification, the similarity search and the corrections of unbalanced reactions in large databases.

When neither the target molecule nor the specific reactions are known in advance, algorithms that classify reactions are useful for library design, such as the *Reaction Class Recommender*, developed by Ghiandoni *et al.* to improve the practical applicability of the predicted routes, [66] and the *Reactions Atlas*, developed by Schwaller *et al.*, to chart the reactions according to similarity. [67]

All those approaches are based on machine learning methodologies, which involve two major aspects: big databases and data accessibility for chemists. Machine learning approaches need large amount of data for training reliable models. Commercial databases of reactions are not as common as databases of molecules. The Reaxys® database, trademark of Elsevier,[68] is one example, as well as the CAS data set. [69] Beside commercial records, a collection of reactions from US patents was made publicly available by Daniel M. Lowe [70] and, recently, the ORD (open-reaction database) was founded, especially for the purpose of supporting machine learning. [71] Such repositories are built on mining scientific articles or patents to extract reaction schemes, conditions and other relevant information. However, the mining process often requires intervention by experts for constant data curation. The accessibility for chemists to algorithms based on machine learning is limited to the functionalities implemented by the developers. In fact, due to lack of background knowledge in programming, chemists might find it challenging to adapt such algorithms to their needs. For this purpose, platforms that save chemists from the burden of programming have been developed, in particular for the purpose of data mining. The most common examples are *RapidMiner* [72] which requires a license and the open-source KNIME [73]. This platform enables scientists with little experience in chemoinformatics to perform calculations and process data according to their research focus. In the DNA-encoded libraries context, a straightforward tool for library design is lacking today.

5.2 Aim

The vastness of chemical reaction space has unfortunately been unexplored in the DEL context, primarily due to the limited applicability of chemical reactions in the presence of a DNA-tag. However, recent advances in DEL technology have expanded the reaction conditions that can be explored, and a tool to select reactions for library synthesis is required. Such tool is demanded to examine large databases of reactions, such as the Reaxys® database, and to apply relevant filters due to idiosyncratic requirements for DEL chemistry such as the oxidation potential of the reagents involved in the reactions, pH, temperature and solvent. Additionally, the algorithm should project the reactions in a three-dimensional plot for visualization of chemical reaction space and uncomplicated interpretation of the data. Finally, it should suggest potential chemistries for DEL synthesis. The whole process must be faster than the manual search and easily accessible to scientists without prior knowledge in chemoinformatics.

5.3 Methods

5.3.1 KNIME workflow

The final KNIME workflow for this work was executed on a machine with the following technical features: memory: 7.6 GB, processor: Intel® Core™ i7-6498DU CPU @ 2.50GHz × 4, graphics: Mesa Intel® HD Graphics 510 (SKL GT1).

The complete scheme of the workflow is depicted in Figure 9 and the five modules are coloured in different hues. Beside a preliminary phase for data preparation, the first two modules filtered the data set according to reaction conditions and to combinatorial significance, respectively. Reactions which featured duplicate reactants or that produced more than one compound were excluded. The third module was dedicated to the description of the reactions which was utilized by the fourth module to arrange them in the three-dimensional chemical reaction space. The fifth module included insights into the final data set and the selection of reactions for application on DNA-tagged substrates.

5.3.2 Input database

The aromatic aldehyde was the elected functional group to start developing the algorithm. This class of building blocks are very versatile and they allow for accessing a plethora of different scaffolds by a large number of reactions. Therefore, they are privileged building blocks for DELs. [74] A sample of 100,000 reactions starting with aldehydes represented the initial data set.

5.3.3 Data preparation

The algorithm started with a data preparation section to select only reactions containing both reactants and products and with commercially available components. Hence, as the Reaxys® data set presents reactions conditions along with the scheme, the database was filtered according to conditions such as temperature, reagents, catalysts and solvents. At first, for standardizing the format, the temperature column was treated with a small workflow. Reactions involving temperatures higher than 200°C were excluded because normally these temperatures are incompatible with DELs, since it leads to DNA denaturation. Although 200°C was a significantly high threshold for reactions involving a DNA-tag, reactions conducted at high temperatures (close to 200° C) were still considered since it could be possible to drive such reactions to completion by increasing the concentration of the reagents at suitable (lower than 100° C) temperatures.

Chemistry selection

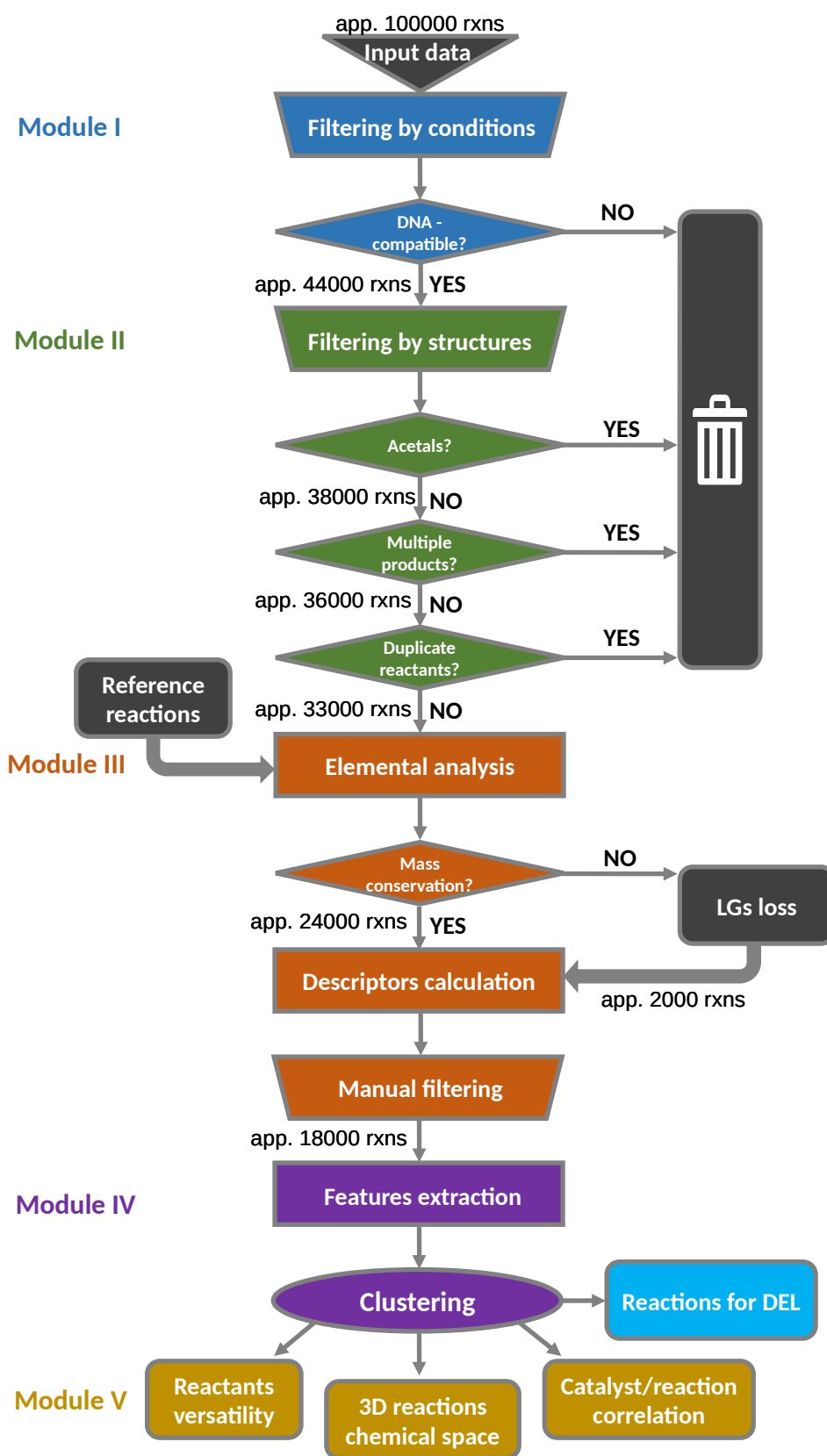


Figure 9: Complete scheme of the KNIME workflow, with its five modules: the first two for filtering, the third for reaction description and the fourth for clustering in the 3D space. The fifth module included all the results of the analysis. rxns =reactions.

5.3.4 Filtering by conditions

The reagents and catalysts were treated together because, in the Reaxys® database, they are not distinctive and were scored according to DNA compatibility. The scoring ranged from zero for incompatible reagents to four for certainly compatible reagents. Reagents that are surely incompatible were identified as mutagens, strong acids and bases, and oxidants. Certainly compatible reagents are represented by neutral buffers or salts, aminoacids or ligands. The intermediate values of 1 and 3 were assigned to reagents that were most likely incompatible and compatible, respectively. Such reagents would show similarity with a reagent classified with a 0 or a 4. The score 2 was assigned to reagents with unknown effect. The scoring was done manually after extracting all the reagents from the database and a complete list of the scored reagents can be found in the *Appendix* as KNIME report. Reagents and catalysts were also scored according to practical applicability, wherein the filtering was tailored for the synthesis on solid support, in our case CPG (controlled pore glass), which carries protected DNA tags that are cleaved under basic conditions. Furthermore, reagents that remain in solid state throughout the reaction, such as resins, were considered not suitable for the synthesis on CPG. Since this strategy allows for usage of organic solvents, they were prioritized according to their boiling points. Solvents with high boiling point are preferred in DEL synthesis, due to the set up in polypropylene tubes that can not be air-tight sealed and long reaction times. With all these factors in mind, the reactions were ranked with the Pareto ranking protocol, a multi-parameter optimization method. [75] In this way, reactions with temperature between 20°C and 80°C, high reagents score and high solvent boiling point received a higher ranking. Reactions with score zero were discarded as they would not be applicable to DNA-encoded chemistry.

5.3.5 Further refinement

The scored reactions were submitted to further filtering. At first, reactions yielding aldehyde or acetal structures were excluded. The former because the reaction occurred on another reactive group of the molecules leaving the aldehyde unreacted and thus, it did not represent an option for our workflow. The latter because acetals are highly susceptible to hydrolysis and usually not desirable in drugs. Reactions producing metal complexes were excluded as well, as such products do not represent viable options in terms of DNA-encoded libraries. Salts, such as hydrochlorides, were removed from the reaction schemes since they would be considered reagents otherwise, impairing future steps of the

workflow. For this purpose, a list of salts extracted from the database was compiled (*Appendix as KNIME report*).

Some filtering criteria were also applied to reactants and products. On one hand, reactions involving two or more identical reactants were discarded because such transformations would not improve the diversity of the final product, only rising the molecular weight. On the other hand, reactions with multiple side products were excluded as well, because, during DEL synthesis a complex mixture is formed making it impossible to separate the side products. Stereoisomers were not considered as separate products since they are not distinguished in the DEL technology.

5.3.6 Reference reactions

In the data set, 33 reactions were inserted manually to function as landmark in the final map. Those entries represented well-known DEL compatible reactions, published in the literature, such as the Ugi, the Povarov or the SnAP reactions. [22], [57], [76] Although such reactions already accounted for examples, their simplest and most general schemes were used as landmarks (*Appendix as schemes*).

5.3.7 Analysis of the elements

After analysing the intermediate results, it became evident that not all reactions presented the same atomic count in reactants and products. Essentially, a part of the reactions was not balanced in mass and this feature, similar to the presence of salt, could impair the calculations of the reaction descriptors. Therefore, each element was counted and the differences between reactants and products for each atom were calculated. Reactions with a null difference were balanced and therefore could access the rest of the workflow untouched. In reactions with a positive difference, mass was added to the products but it did not appear among the reactants in the reaction scheme. In fact, in such reactions, some reactants were missing or not correctly represented due to data management errors. Violating the law of mass conservation, those entries were not corrected in any way and discarded. In contrast, reactions with a negative difference showed the presence of leaving groups (LGs), i.e. parts of the reactants there were lost during the reactions. Common examples of this group of reactions are nucleophilic substitutions or the Suzuki coupling. To correct those entries, a list of LGs was compiled (*Appendix as KNIME report*) based on the data and the LGs were matched with the reactants which contained them as substructures. While matching the LGs, the atoms of the LG were subtracted from the respective reactant prior to the descriptors calculation, within a parallel workflow.

5.3.8 Reaction descriptors calculation

The balanced reactions were then described according to molecular descriptors inspired by the work of Feher *et al.* [77] The 21 features listed in Table 1 were chosen among others because they accounted for substructures contained in the molecules involved in the reactions. With these descriptors, the focus was placed on the changes happening during the reactions, in order to group them by similarity and map the chemical reaction space. For this reason, after counting the chemical substructures present in the products and the reactants, they were subtracted. In this way, only the information about the modifications was stored, neglecting other moieties which did not participate directly in the reaction, such as substituents or peripheral groups. The idea of calculating reaction descriptors as difference was already described in literature by Schneider *et al.* [78]

Table 1: Descriptors with respective type in brackets and meaning.

Descriptor (type)	Meaning
NumRings (Number (integer))	number of rings
NumAromaticRings (Number (integer))	number of aromatic rings
NumAliphaticRings (Number (integer))	number of aliphatic rings
NumAromaticHeterocycles (Number (integer))	number of aromatic heterocycles
NumAliphaticHeterocycles (Number (integer))	number of aliphatic heterocycles
NumAromaticCarbocycles (Number (integer))	number of aromatic carbocycles
NumAliphaticCarbocycles (Number (integer))	number of aliphatic carbocycles
NumRotatableBonds (Number (integer))	number of rotatable bonds
NumAmideBonds (Number (integer))	number of amide bonds
NumAliphaticBonds (Number (integer))	number of aliphatic bonds
NumAromaticBonds (Number (integer))	number of aromatic bonds
NumCisTransBonds (Number (integer))	number of cis/trans bonds
Csp3 (Number (integer))	number of sp ³ carbons
C-C (Number (integer))	number of carbon-carbon bonds
C-N (Number (integer))	number of carbon-nitrogen bonds
C-O (Number (integer))	number of carbon-oxygen
C-S (Number (integer))	number of carbon-sulfur
NumNInR (Number (integer))	number of nitrogen in rings
NumOInR (Number (integer))	number of oxygen in rings
NumPInR (Number (integer))	number of phosphorus in rings
NumSInR (Number (integer))	number of sulfur in rings

5.3.9 Feature extraction

The 21 features characterizing the reactions represented the dimensions of the chemical reaction space but, as they were, it was hard to interpret them and impossible to visualize them in the three-

dimensional space. Therefore, they were summed forming three variables, as shown in Figure 10. In particular, the features related to cyclic substructures were grouped together after performing a correlation filter. This filter avoided redundant information which could bias the clustering and it was necessary because some correlations were noted among the 21 features (Figure 11). Redundancy of an information could affect the clustering and further analysis because that information would be considered more important than others. Correlations can be positive, if two variables' values grow or decline proportionally, or negative, if growth of one variable correspond to decline of the second one. For example, it was worth noticing the negative correlation between the number of C-O bonds and the number of C-N bonds, typical of substitution or condensation reactions. This finding would mean that considering only the number of C-O bonds or the number of C-N bonds would be sufficient to describe the reactions. The same procedure was applied to features related to bonds and to heteroatoms. At the end, three variables described the reactions, *rings*, *bonds* and *heteroatoms*, and they represented the dimensions of the chemical reaction space.

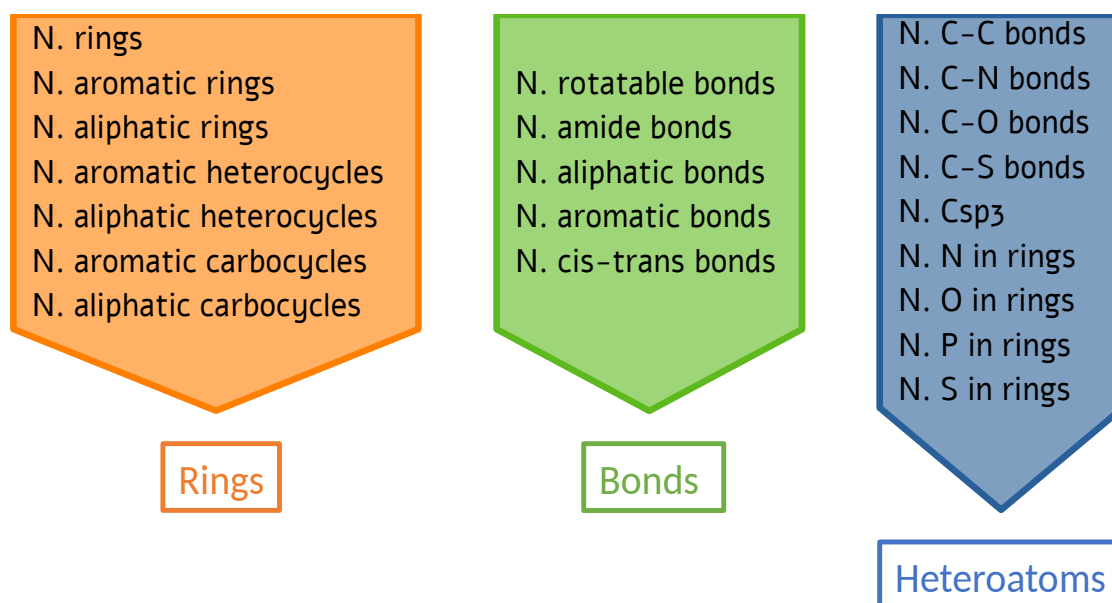


Figure 10: Grouping of the 21 features into the three new variables.

Chemistry selection

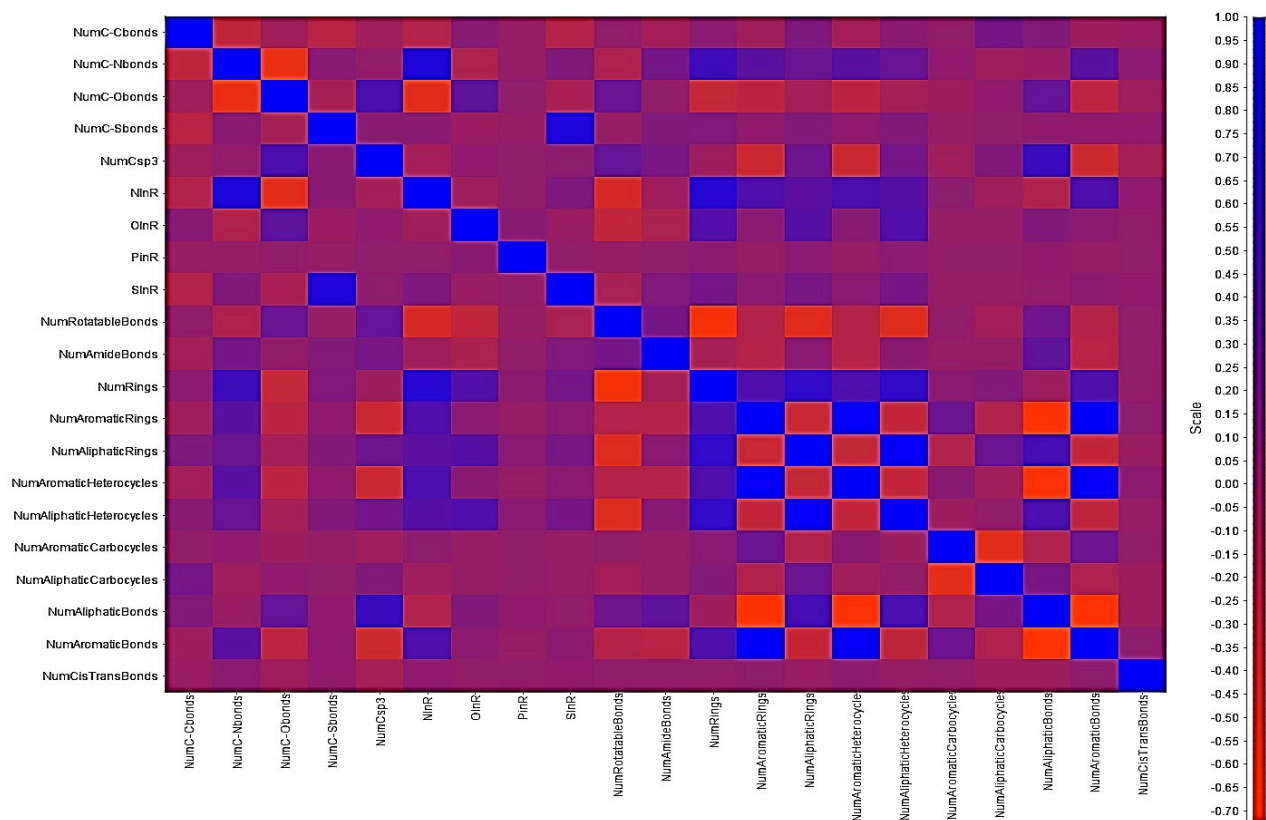


Figure 11: Correlation matrix highlighting the correlation between features. Deep blue and deep red characterize strong positive and negative correlation, respectively.

5.3.10 Reaction classification

Notably, the three variables could assume positive or negative values if the described substructure was formed or converted during the reaction. For example, if one ring was closed the variable *rings* would assume a value of 1, whereas if one ring was opened the same variable would assume the value of -1. This distinction between forming (above zero) or breaking (below zero) variables was employed for classifying the reactions according to hard rules (Table 2).

Table 2: Hard rules and corresponding classes for the reactions classification.

Rule	Class
\$BONDS\$ = "breaking" AND \$RINGS\$ = "forming" AND \$HETEROATOMS\$ = "forming" => TRUE	Multi-rings or N-heterocycles formation
\$BONDS\$ = "breaking" AND \$RINGS\$ = "breaking" AND \$HETEROATOMS\$ = "breaking" => TRUE	Aldol condensation, ring opening/closure
\$BONDS\$ = "forming" AND \$RINGS\$ = "breaking" AND \$HETEROATOMS\$ = "breaking" => TRUE	Aldol condensation, ring opening/closure
\$BONDS\$ = "breaking" AND \$RINGS\$ = "forming" AND \$HETEROATOMS\$ = "breaking" => TRUE	Ring formation
\$BONDS\$ = "forming" AND \$RINGS\$ = "forming" AND \$HETEROATOMS\$ = "breaking" => TRUE	Ring formation
\$BONDS\$ = "forming" AND \$RINGS\$ = "forming" AND \$HETEROATOMS\$ = "forming" => TRUE	Heterocycles formation
\$BONDS\$ = "forming" AND \$RINGS\$ = "breaking" AND \$HETEROATOMS\$ = "forming" => TRUE	Grignard reaction, cyanation, aminoalkylation
\$BONDS\$ = "breaking" AND \$RINGS\$ = "breaking" AND \$HETEROATOMS\$ = "forming" => TRUE	Aminoalkylation

5.3.11 Clustering

The three variables were employed as attributes to apply the clustering protocol. The "fuzzy" c-means clustering algorithm was chosen because it has been proved that it performs better than other algorithms with sparse data sets. [79] In fact, even after normalization, the variables could assume integer values, since they described substructures that could be present in total or not at all. The cluster quality was assessed by the *silhouette* coefficient, calculated as:

$$Sc = \frac{(b-a)}{\max(a,b)} \quad (\text{Eq. 1})$$

The terms b and a are referred to as separation coefficient (inter-cluster distance) and cohesion coefficient (intra-cluster distance), respectively. The Equation 1 implies that the more separated (big b) the clusters are among each other and the more cohesive (small a) they are within each other, the better is the clustering quality. Before using the *silhouette* coefficient as measure of the cluster quality, its correlation with a chemically sound clustering was assessed. For this purpose, the data set was sampled by extracting central and peripheral entities (or reactions) from each cluster and the sample was validated by visual inspection. The same visual inspection permitted the identification of clusters of interest, that could be investigated deeply to select the reactions to be translated into DNA-encoded chemistry. A short sub-workflow was designed to optimize the clustering parameters "number of cluster" and "lambda". The lambda parameters refers to the shape of the clusters and to improve the clustering it must be adapted to the data set. [80] With the optimized conditions, the clustering was performed and important information was extracted from the data set, as shown in the following sections.

5.3.12 Reactants and catalysts versatility

In order to prioritize reactants and catalysts for their use in DNA-encoded chemistry, the versatility of both was analysed. The term versatility in this case signified the occurrence of the considered reactants or catalysts within the data set, especially within many distinct clusters. Considering that each cluster corresponded to at least one reaction type, if a reactant was part of many clusters, it could potentially be used in library synthesis to build different scaffolds, greatly improving the diversity of the final library. On the other hand, catalysts promoting many types of reactions could be worth screening in the DEL context, to assess their DNA compatibility. Since the DEL chemical space is limited by the lack of DNA compatibility of metal catalysts, this information has the potential to expand the

DEL toolbox. In particular, in our group the damaging effect of metal catalysts has been investigated and it has been proved that metal catalysts are more problematic than, for example, organocatalysts, especially in combination with forcing reaction conditions. [22] Therefore, their versatility was investigated at first. Additionally, for the same reason, the correlation between the clusters and the metal catalysts was explored. Catalysts promoting attractive reaction types would be prioritized in the screening experiments.

5.3.13 Scaffolds in drugs

To analyse the data set in terms of accessible drug-like scaffolds, 15 of the most commonly found rings were compiled and searched throughout the data set as substructures. This allowed us to uncover the landscape of accessible products from the aldehyde building block.

5.3.14 Anti-reactions

One group of reactions attracted our attention because it involved nucleosides. This characteristic renders such reactions not practical in a DEL context as they could potentially occur on the DNA tag, modifying it and hindering the sequencing step. Therefore, we extracted and excluded all reactions involving nucleosides by substructure search.

5.4 Results and discussion

The set of workflows developed in this work formed an algorithm which from a database of chemical reactions extracted potential candidates for library design. Additionally, the algorithm provided important insights into the available reactions starting from the aldehyde building blocks. At first, the reactions were classified according to the signs of the value of the three variables using hard rules: breaking (-1) or forming (1) of substructures such as bonds and rings. Although by the combination of values (positive and negative) and the variables (*rings*, *bonds*, *heteroatoms*) one would expect eight classes (2^3), after visual inspection, four out of the eight classes were merged into two making a total of 6 classes. An analysis of the classes proportion throughout the data set is represented by the pie chart in Figure 44. Notably, aldol condensations and reactions yielding linear scaffolds dominated in this data set with roughly 10,000 reactions out of 18,000. They are followed by ring-forming reactions accounting for approximately 7,000 reactions.

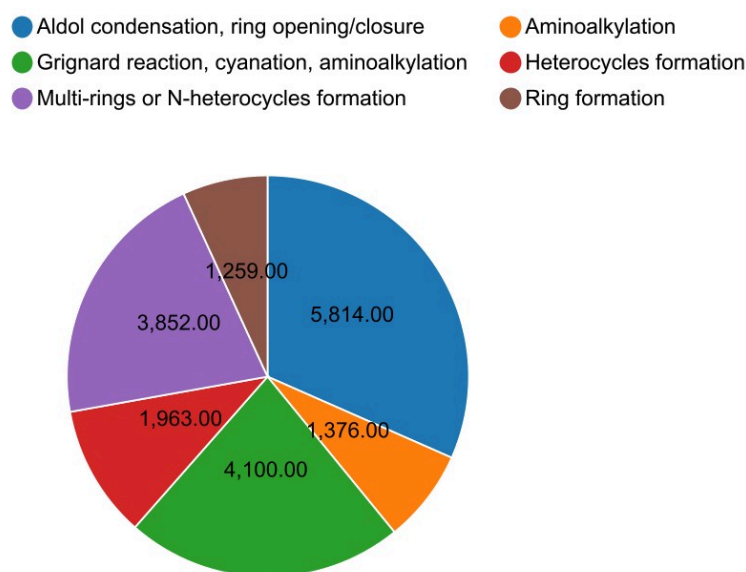


Figure 12: Pie chart for the shares of the reactions classes in the provided data set.

Despite the simplicity of this approach and the hard rules, this classification successfully covered most of the reactions. However, this methodology could be implemented with machine learning approaches, similar to the work of Schwaller *et al.* [67]

Since the three variables represented the dimensions of the 3D chemical reaction space, they divided it in sections as depicted by the scatter plot in Figure 13. The additional lines in the centre correspond to the null in all axes, so they split them by the sign, positive on one side and negative on the other side. Additionally, in Figure 13, two reactions are depicted as examples. The one in the blue box, is an

example of aldol condensation and it is placed in the negative part of the dimension *rings* but in the positive part of the dimension *bonds*, meaning that no rings are formed but only bonds. In this case, the presence of the leaving group did not influenced the projection of the reaction, as unbalanced reactions were priorly processed and corrected. [81] The one in the purple box, represents reactions forming heterocycles and/or more than one ring. In fact, it is placed in the positive area for both variables *rings* and *heteroatoms*, meaning that both substructures are formed during the reaction.[82]

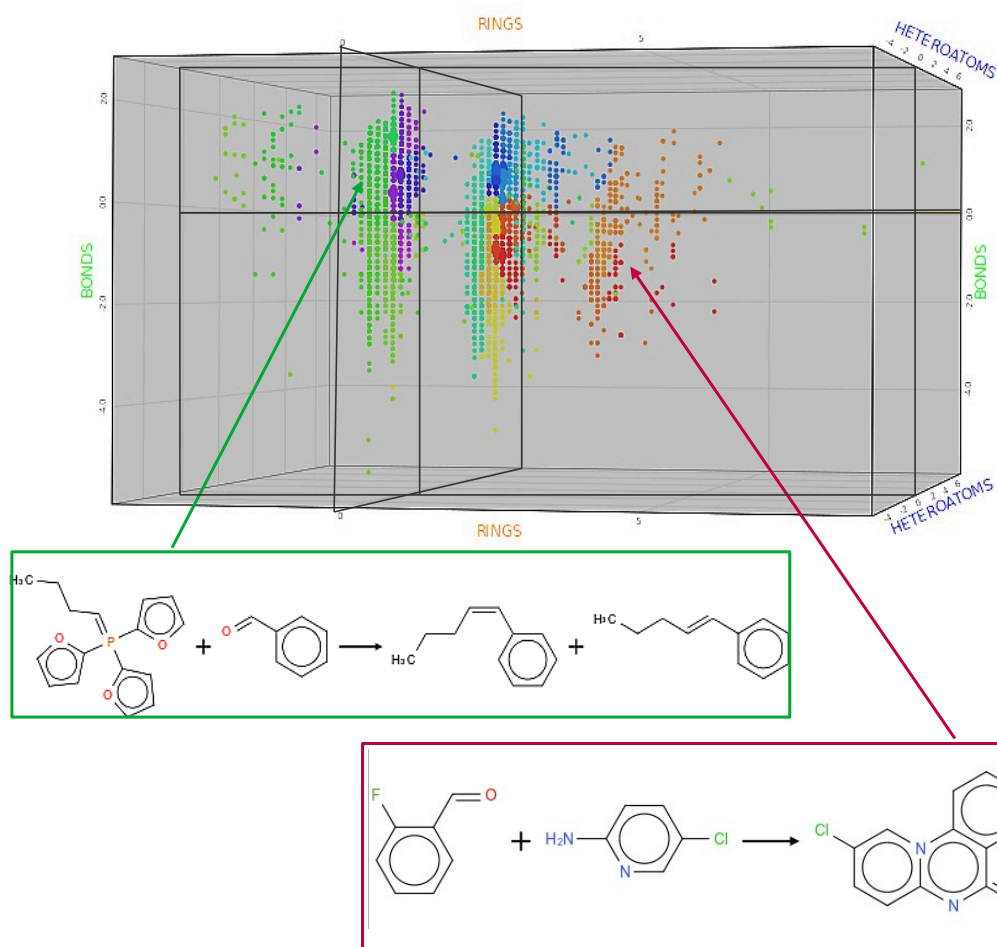


Figure 13: Scatter plot of the chemical space defined by the three dimensions *rings*, *bonds* and *heteroatoms*. The data point are reactions and they are colored according to the clusters. The lines correspond to the value zero in each dimension, dividing the plot in sections. Each section correspond to one class for a total of six classes.

In Figure 14, the data points in the scatter plot representing the chemical reaction space are coloured according to their affiliation to the clusters. The integral nature of the variables emerged in the geometrical planes visible in the scatter plot. However, each geometrical plane was divided into different clusters according to the similarity among the reactions. The data points with bigger size in Figure 14 represent the reference reactions which were inserted manually in the database to function

as landmarks in the reaction landscape. The descriptors calculation and the clustering quality could already be confirmed by the fact that the Ugi and the Cushman reactions belonged to very close clusters, compared to the Petasis reaction which was projected farther away.

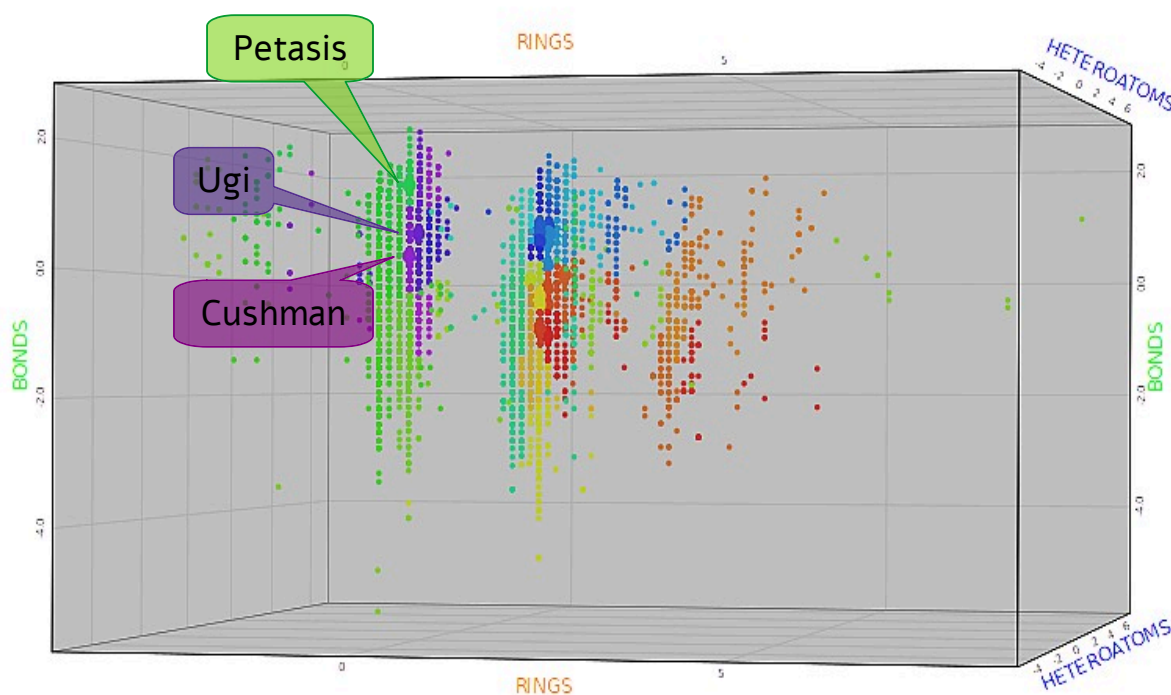


Figure 14: Scatter plot of the clustered reactions within the chemical reaction space defined by the three variables rings, bonds and heteroatoms. Reactions are colored according to the cluster affiliation. The big data points represent the reference reactions, in particular the Ugi (indigo), Cushman (purple) and Petasis (green) reactions.

The quality of the clustering was also assessed from a statistical point of view by the Silhouette coefficient, used for optimization of the clustering parameters such as number of clusters and lambda. With the optimal parameters, the clustering protocol produced 58 clusters and 45 unclustered reactions (Table 3).

Table 3: Summary of all clusters with the respective number of reactions.

Cluster	number of reactions	Cluster	number of reactions	Cluster	number of reactions	Cluster	number of reactions
cluster_29	1520	cluster_53	383	cluster_28	166	cluster_13	72
cluster_48	1455	cluster_51	375	cluster_9	149	cluster_41	67
cluster_52	1195	cluster_4	362	cluster_20	118	cluster_46	67
cluster_11	1046	cluster_50	338	cluster_23	117	cluster_1	48
cluster_47	1045	cluster_26	324	cluster_7	117	cluster_36	48
cluster_37	915	cluster_43	319	cluster_12	113	NoiseCluster	45
cluster_24	877	cluster_55	319	cluster_44	113	cluster_0	45
cluster_56	816	cluster_10	266	cluster_21	110	cluster_39	39
cluster_45	557	cluster_35	248	cluster_40	110	cluster_31	30
cluster_54	557	cluster_33	218	cluster_42	110	cluster_38	29
cluster_5	553	cluster_34	218	cluster_32	92	cluster_49	26
cluster_57	526	cluster_14	204	cluster_19	86	cluster_2	23
cluster_27	448	cluster_22	189	cluster_17	83	cluster_8	23
cluster_30	443	cluster_18	178	cluster_25	80	cluster_6	19
cluster_15	394	cluster_3	170	cluster_16	75		

The largest clusters, with over thousand reactions per cluster, show aldol condensations or reactions producing scaffolds from a linear bond formation, confirming the reaction classification.

Although the noise cluster was highly assorted, it contained some important reactions such as one involving a ring opening reaction (Figure 15A) and a reaction forming sp^3 rich scaffolds (Figure 15 B).

These two examples present DEL-relevant reactions, revealing that the noise cluster cannot be neglected. [83], [84]

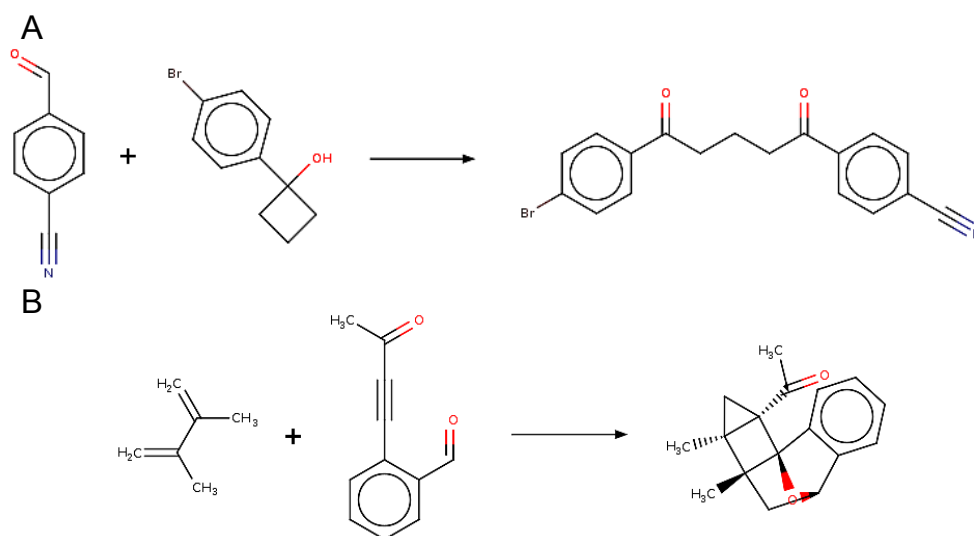


Figure 15: Examples of unclustered reactions from the *noise*: (A) ring-opening reaction [83] and (B) reaction forming an sp^3 -rich scaffold.[84]

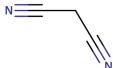
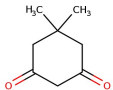
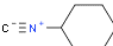

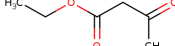
5.4.1 Additional findings

Beside confirming the quality of the clustering and extracting potential reactions for DELs, the algorithm provided further insights into the database, in relationship to the aldehyde functional group.

5.4.1.1 Versatile reactants

The most common reactants excluding aldehydes and amines presented the malononitrile which took part in 552 reactions from 41 clusters, and dimethylcyclohexane-1,3-dione, reacting in 266 reactions out of 38 clusters (Table 4).

Table 4: Top five versatile reactants, excluding aldehydes and amines.

Reactant	Cluster count	RXNs count
	41	552
	38	266
	32	205
	22	167
	26	162

These two reactants were involved in multi-component reactions forming attractive heterocyclic scaffolds (Figure 16).

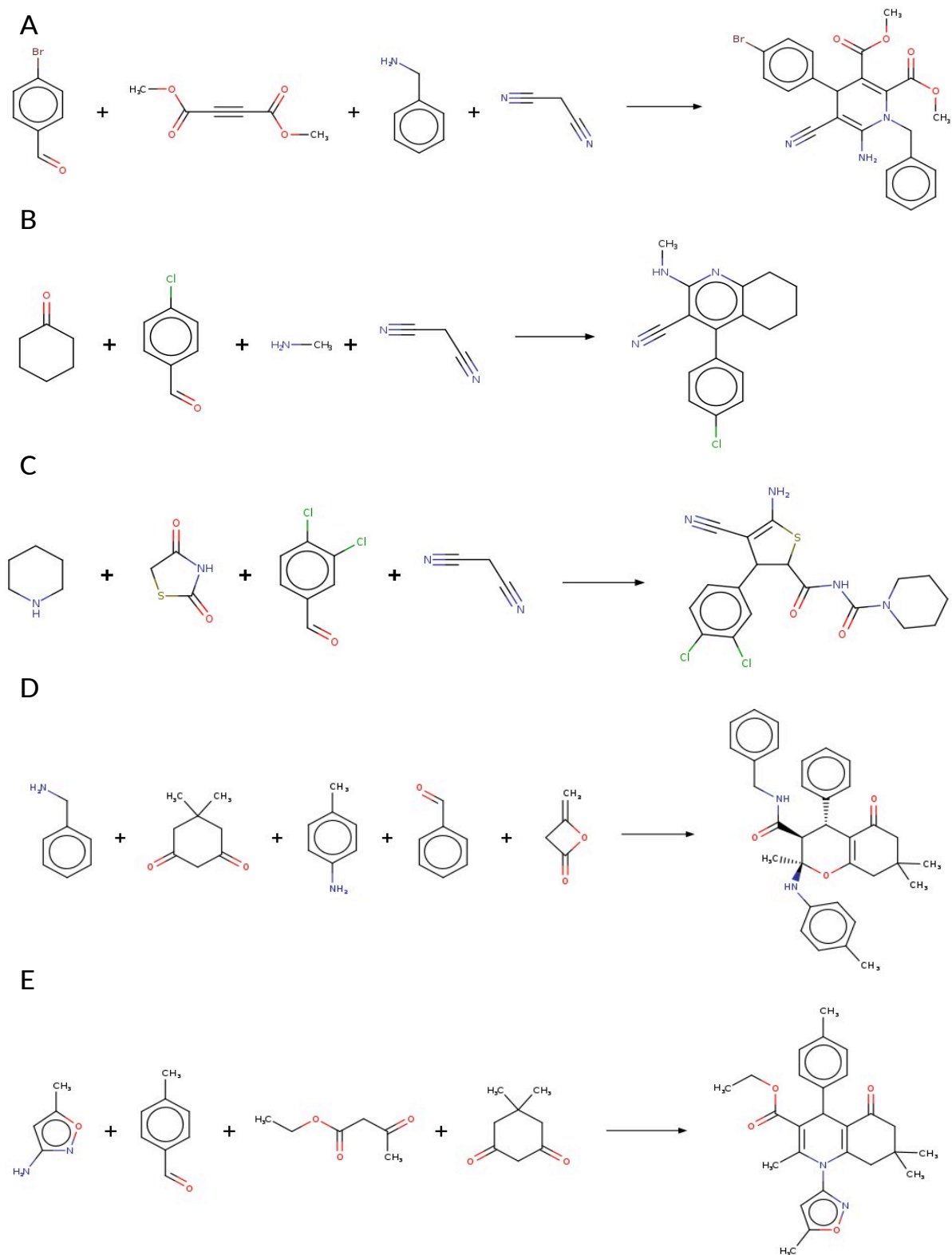


Figure 16: Example reactions with (A), (B), (C) malononitrile and with (D), (E) dimethylcyclohexane-1,3-dione. [85], [86], [87], [88], [89]

5.4.1.2 Metal catalysts analysis

These results were uncovered in collaboration with the bachelor students in the Brunschweiler group, Lars Grützbach and Felix Biesenkamp. At first, all the metal centres, i.e. metals as part of salts or coordinated by ligands, were extracted manually from the data set and the respective reactions counted. It is worth noticing that almost 10,000 reactions did not report any metal catalyst in the data set, which accounted for about 70 % of the entries. In such reactions, either no catalyst or reagent was reported at all in the initial table or the reported catalyst did not contain a metal centre. In the first scenario, only a more accurate report of data in scientific publications might solve the issue (except for reactions that occur without catalyst), whereas in the second scenario, including organocatalysts would improve the performance of the analysis. The most often utilized catalyst in this database was zinc, promoting 365 reactions, followed by titanium with 280 entries. The complete list of metal centres, with the respective counts is provided in Table 5.

Furthermore, a correlation between the metal centres and the reaction type was investigated. Since the quality assessment of the clustering proved that each cluster featured the same reaction type or very similar ones, the relationship between the clusters and the metal centres was explored (Figure 17).

Table 5: Metal centres with respective frequencies.

Query metal centre	Occurency count	Query metal centre	Occurency count
missing	8348	[As]	8
undefined	1712	[Cr]	7
[Zn]	365	[Ga]	7
[Ti]	280	[Mn]	7
[Cu]	202	[Ru]	6
[Ag]	197	[Hf]	4
[Pd]	181	[Pb]	4
[Al]	97	[V]	4
[Yb]	94	[Y]	4
[Fe]	82	[Mo]	3
[Sn]	77	[Nd]	3
[Sb]	54	[Pt]	3
[Ce]	52	[Lu]	2
[Bi]	40	[Pr]	2
[In]	25	[Re]	2
[Ni]	21	[Te]	2
[Zr]	21	[Er]	1
[Nb]	18	[Hg]	1
[W]	14	[Sm]	1
[Au]	12	[Ta]	1
[La]	10	[U]	1
[Co]	9		

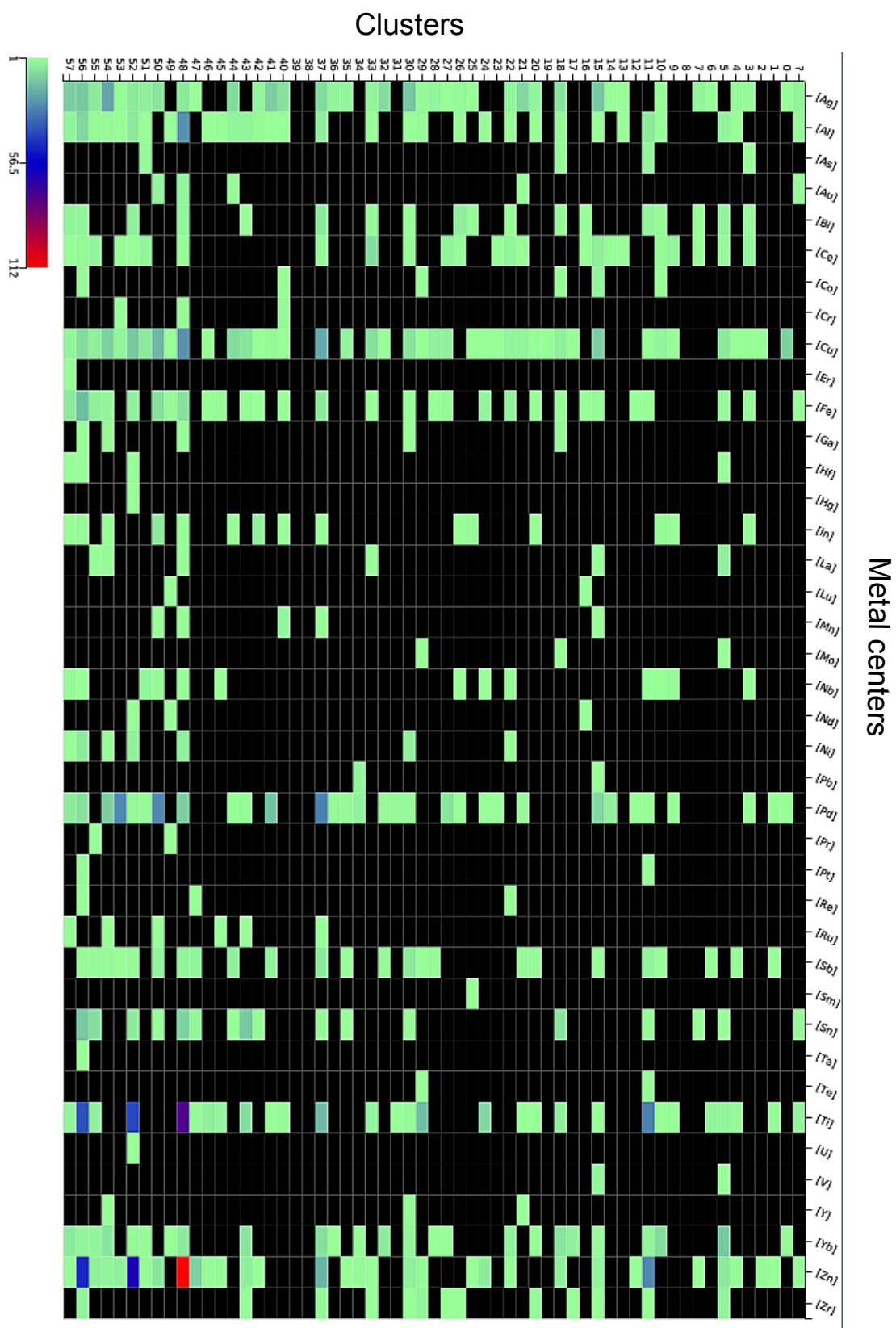


Figure 17: Heat map of the correlation between metal centres (y-axis) and clusters (x-axis). The cells are colored according to the number of reactions as illustrated by the scale legend.

Zinc and titanium proved to catalyse mostly aldol condensations or A³-couplings such as the ones depicted in Figure 18. [90], [91]

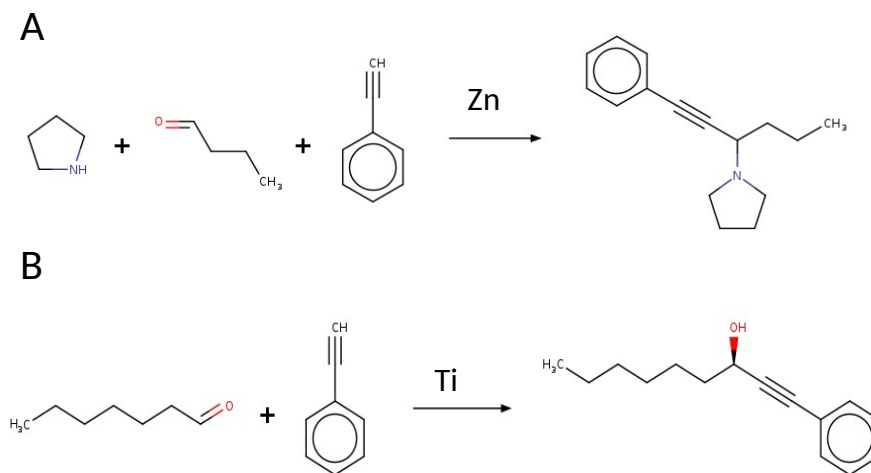
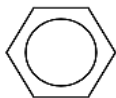
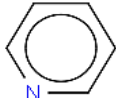
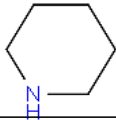
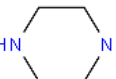
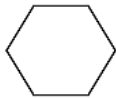
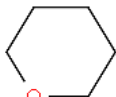
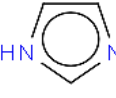
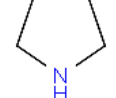

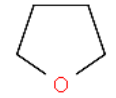
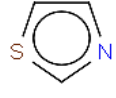
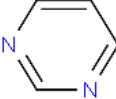


Figure 18: Example reactions: (A) A³-coupling catalysed by the metal zinc and (B) addition catalysed by the metal titanium.

5.4.1.3 Accessible scaffolds

Finally, the accessibility of 15 most common scaffolds in drugs structures was investigated and the results of the analysis are listed in Table 6. Although Table 6 is sorted according to the frequency in the initial publication,[20] it is noticeable that tetrahydropyran accounted for 34 reactions, followed by the pyridine with 19 reactions.

Table 6: Table of accessible scaffolds. The reference count refers to the publications reporting the most common rings in drugs. [20]

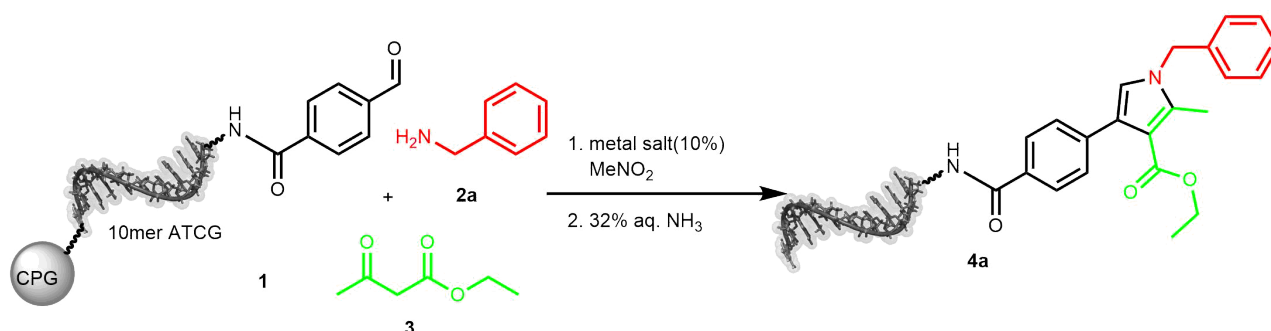
Scaffold	Cluster count	RXNs count	Reference count
	7	8	538
	8	19	54
	3	7	54
	1	1	51
	7	17	38
	8	34	32
	6	10	30
	7	14	29
	2	3	28
	5	11	27
	6	15	25
	5	7	20

5.4.2 Experimental validation of applicable reactions

After performing the clustering, each cluster was sampled by extracting five reactions close to the cluster centre and five reactions at the periphery of the cluster. This process allowed for visually inspecting the representative reactions for each cluster and select the most significant in the context of library synthesis. Two reactions were selected and tested on DNA-tagged substrates for their compatibility with the DEL process.

5.4.2.1 Pyrrole synthesis starting with aldehyde, amine and ethylacetoacetate

This reaction was selected from a cluster of multicomponent reactions for the formation of the pyrrole scaffold, which is common in drugs. [20] It fulfilled the requirements of potential reactions for DNA-encoded chemistry, being a multi-component reaction and showing a large scope of substrates. Although in the original publication the reaction temperature was set to the boiling point of nitromethane (103°C), it could be applied at lower temperature on DNA, due to the excess of the other reagents (Scheme 1). [92] The results in terms of conversion and DNA degradation are shown in Table 7. An increase in temperature to 80 °C was found to be necessary to drive towards product formation since 60 °C proved inadequate, independent of the metal catalyst. However, at 80°C, a reaction time longer than 6 hours caused degradation of the DNA barcode, reducing the conversion. The optimal ratio between conversion and degradation was reached at 80°C for 6 hours using the metal salt FeCl₃ with 25% conversion and 30% degradation. In contrast, with the metal salt NiCl₂ conversion and degradation accounted for 19% and 52% respectively. The high value of degradation could be attributed to the formation of hydrochloric acid during the reaction. However, in the HPLC trace in Figure 19, it is shown that the product can be easily separated from the starting material and eventual degradation products, so the reaction can be used for library synthesis.



Scheme 1: Pyrrole synthesis on DNA-aldehyde conjugate.

Table 7: Optimization of pyrrole synthesis on a CPG-bound DNA-aldehyde conjugate **1**.^a

Entry	Metal salt	T (°C)	Time (h)	Conversion [%] ^b	Degradation [%] ^c
1	FeCl ₃	60	6	0	20
2	FeCl ₃	60	8	N.d.	N.d.
3	FeCl ₃	80	6	25	30
4	FeCl ₃	80	8	14	45
5	NiCl ₂	60	6	4	25
6	NiCl ₂	60	8	N.d.	N.d.
7	NiCl ₂	80	6	19	52
8	NiCl ₂	80	8	11	27

^a The suspension of DNA-aldehyde conjugate **1** (20 nmol), benzylamine **2a** (250 μM), ethylcetoacetate **3** (250 μM) and metal salt (25 μM) in solvent (40 μL) was shaken at the given temperature for the given time. DNA cleavage with 32% aq. Ammonia solution at ambient temperature for 4 h. ^b Determined by RP-HPLC analysis based on the ratios of **1** to **4a** (AUC). ^c Determined by RP-HPLC analysis and scaled to the purity of the CPG-coupled DNA-aldehyde conjugate **1**. MeNO₂ = nitromethane. N.d. = not detected, 10mer ATGC = 5'-NH₂-C₆-GTCATGATCT-3'-CPG.

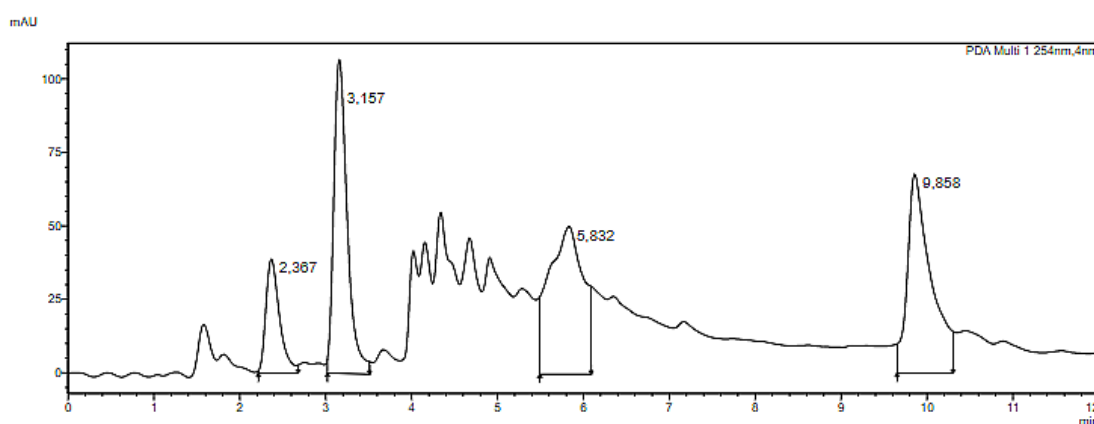
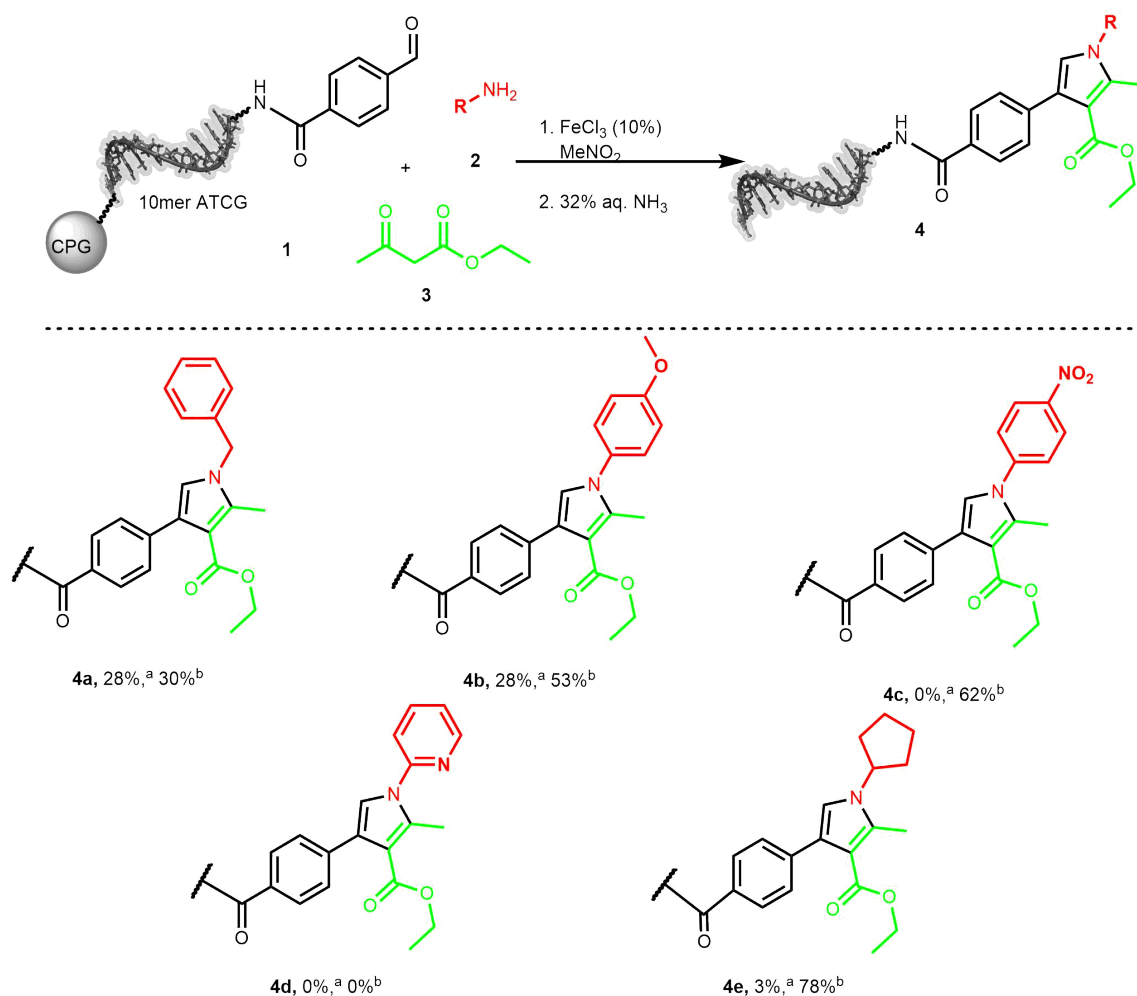


Figure 19: HPLC trace of the crude mixture after the pyrrole synthesis performed at 80°C for 6 hours with FeCl₃. The peaks at 2.3 and 3.1 minutes are degradation products (37% area), while the peak at 5.8 minutes is the starting material (32% area). The exact masses of the materials were measured by MALDI-TOF, whose spectra are reported in Table 22 of the Experimental part. Product is detected at 9.8 minutes (30% area).

Using the optimized conditions, a scope of amines was investigated to assess the applicability of this reactions in the combinatorial context. A set of substrates, which shows diverse substitutions pattern and reactivity. Therefore, benzylamine, *p*-methoxyaniline, *p*-nitroaniline, 2-aminopyridine and cyclopentylamine were reacted with the DNA-aldehyde conjugate **1** under the optimized conditions (Scheme 2 and Table 8). The benzylamine, being the best performing substrate in the original

publication, [92] showed the highest conversion of 28%. The *p*-methoxyaniline reacted equally well, most likely due to the activating methoxy group. In contrast, the *p*-nitroaniline did not perform well, not delivering the product, due to the deactivating nitro group in *para* position to the amino group. Finally, 2-aminopyridine did not form the desired product and cyclopentylamine yielded only traces of product with high degradation. The reason for the degradation with the last amine could only be connected with the reactivity of the amine, since the metal salt concentration and the temperature were maintained the same as the other experiments. Overall, the reaction delivered the desired product, it was applied to diverse starting materials and the respective products were detected according to their reactivity.



Scheme 2: Scope of amines for the pyrrole synthesis on the DNA-aldehyde conjugate.

Table 8: Amine scope for pyrrole synthesis on a CPG-bound DNA-aldehyde conjugate **1**.^a

Entry	Product 4	Amine 2	Conversion [%] ^b	Degradation [%] ^c
1	4a	2a	28	30
2	4b	2b	28	53
3	4c	2c	0	62
4	4d	2d	0	0
5	4e	2e	3	78

^a The suspension of DNA-aldehyde conjugate **1** (20 nmol), amine **2** (250 μ M), ethylacetoacetate **3** (250 μ M) and FeCl₃ (25 μ M) in MeNO₂ (40 μ L) was shaken at 80 °C for 6 h. DNA cleavage with 32% aq. Ammonia solution at ambient temperature for 4 h. ^b Determined by RP-HPLC analysis based on the ratios of **1** to **4**. ^c Determined by RP-HPLC analysis in ratio to the purity of the CPG-coupled DNA-aldehyde conjugate **1**. MeNO₂ = nitromethane. 10mer ATGC = 5'-NH₂-C6-GTCATGATCT-3'-CPG.

5.4.2.2 Pyrrolidine synthesis starting with aldehyde, aniline and thioester

This reactions produced an attractive scaffold as well, rich in Csp³ which increases the three-dimensionality of the structure. Interestingly, the reaction undergoes a ring expansion to form a 5-membered pyrrolidine ring starting from a 3-membered cyclopropane ring. It proceeded at room temperature constituting a valid option for DNA-encoded chemistry. [93] The reaction was firstly tested on a pyrimidine-based DNA under the conditions from the original publication, adapted to DNA-encoded chemistry (Scheme 3). The results of these experiments are summarized in Table 9. Despite two metal salts being explored for this reaction, none of them enabled the formation of product.

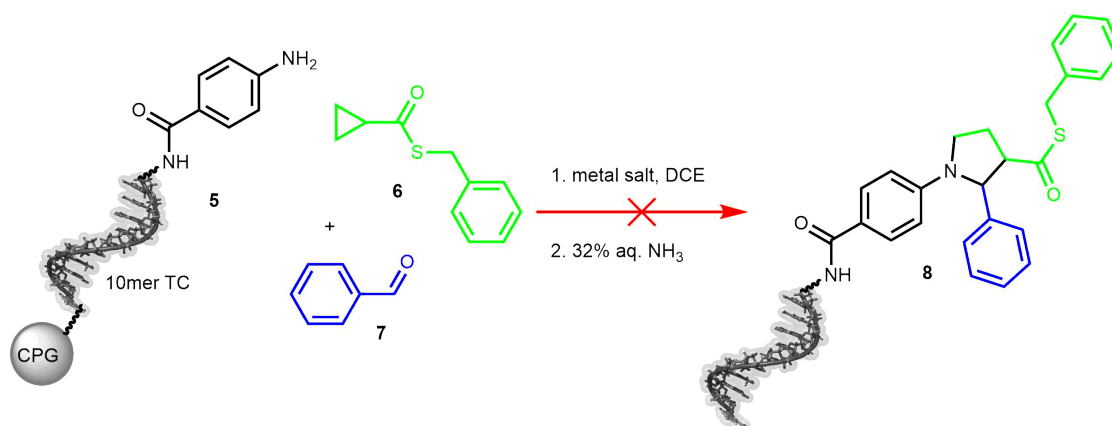
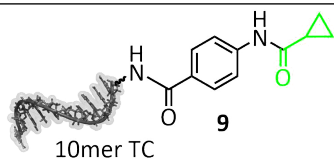

**Scheme 3:** Test reaction for pyrrolidine synthesis on DNA-amine conjugate.

Table 9: Test reaction for pyrrolidine synthesis on a CPG-bound pyrimidine DNA-amine conjugate **5**.^a

Entry	Metal salt	Conversion ^b	Identified side product ^b
1	Et ₂ AlI	n.d.	 10mer TC 9
2	MgI ₂	n.d.	 10mer TC 10

^a The suspension of DNA-amine conjugate **5** (20 nmol), thioester **6** (200 mM) and benzaldehyde **7** (200 mM) and metal salt (200 mM) in DCE (40 μ L) was shaken at ambient temperature for 1 h. DNA cleavage with 32% aq. ammonia solution at ambient temperature for 30 min. ^b Determined by MALDI-TOF analysis, DCE = dichloroethane. n.d. = not detected, 10mer TC = 5'-TTC CTC TCC T-3'-CPG.

Instead, two side products could be detected by MALDI-TOF analysis (compound **9** and **10**). The former was formed by the reaction between the amino group of the DNA-aniline conjugate and the thioester, meaning that the nucleophilic substitution was faster than the cyclization. The latter derived from the cleavage of the C6-amino linker of the DNA barcode, caused by the metal salt MgI₂. [94] This finding rendered this metal salt unsuitable for DELs and might help in improving the filtering cascade.

5.5 Conclusions

In this chapter, the design of a novel algorithm to select reactions for DNA-encoded libraries is reported. Such reactions are extracted from large chemistry databases and must meet requirements linked to compatibility with DNA and with the combinatorial process. The elected functional group to start the search was the aldehyde for its ample usage in medicinal chemistry. [74] To adapt the reactions to DNA-encoded chemistry, the algorithm included an initial filtering cascade which removed reactions involving harsh conditions, followed by a filtering step to optimize the efficacy in the combinatorial context. An original set of descriptors was utilized to describe the reactions and project them in the chemical reaction space. Such descriptors revealed their effectiveness in discerning similar and dissimilar reactions projecting them in different areas of the chemical space accordingly. Furthermore, the reactions were clustered and the clusters sampled in order to reduce to a minimum the data to be manually inspected. Attractive reactions for library design were extracted from the clusters and translated onto DNA-conjugates: a multicomponent pyrrole synthesis and a pyrrolidine synthesis by ring expansion. The former successfully delivered product, whereas the latter uncovered a metal salt which degraded the DNA code before the reaction could take place. This information might result essential for improving the filtering cascade and it underlined the need to generate more experimental data about DNA-compatible reagents. Additionally, insightful information could be gained about versatile reactants, accessible scaffolds from the aldehyde building block and about the most used catalysts or reagents. In summary, the algorithm provided insightful information about a data set of reactions starting with aldehydes. Such information could not have been retrieved manually and in a short time. In fact, the whole analysis progresses over approximately two-three hours depending on the CPU. Although not necessary, the human intervention can be implemented at different stages of the workflow. For example, the filtering cascade can be completely customized according to other contexts, such as peptides chemistry. Modifying the descriptors and the clustering method is not advisable, but it could become necessary when using very different data as input. Finally, further information can be extracted from the data depending on the research focus. The whole algorithm was developed in KNIME, especially for this reason: allowing chemists, lacking prior experience in programming, to operate it and adapt it to their needs. Over the course of the chapter, some limitations for the generalization of the algorithm emerged, such as the manual enumeration of leaving groups or the reactions classification. These two limitations might be overcome by employing

machine learning techniques. The leaving group enumeration could be improved using the current results as training data and predicting eventual leaving groups present in other data sets. The reactions classification could be translated into something similar to the *Reaction Atlas* developed by Schwaller *et al.* [67] Lastly, data curation remains the most important challenge in this field, as good models derive from good data. Great attention is to be put in mining scientific publications or lab notebooks, to extract complete information that can be processed by chemoinformatics methodologies, and to include all those information in publicly available databases. [66], [71]

6 Building blocks selection

6.1 Introduction

The design of DNA-encoded libraries is oriented in two directions. The first is represented by focused libraries that are designed for a specific target. A focused library contains specific features and known motives in order to bind to the selected target. The second direction is represented by libraries that are not target specific. In this second case, library members are as diverse as possible in order to cover a larger chemical properties space. While in the first case the chance of success is higher due to established binding modes, in the second case *hits* might show completely novel binding modes and unexpected features. The unpredictability of the *hit matter* for diverse libraries renders them riskier but the novelty in the case of *hit* identification is higher.

6.1.1 Chemical space coverage

Reported literature on DELs suggests that the chemical diversity of library members has a bigger impact on DEL productivity than other factors such as library size. In fact, very large libraries characterized by low scaffold diversity have a lower *hit* identification rate than smaller but structurally diverse libraries.[95] To define diversity, though, a premise about chemical space is to be made. Chemical space is defined as the ensemble of all feasible chemical entities, and the properties space can be described as the chemical space delimited by specific properties. [96] Diversity in the library context can be defined as extensive coverage of chemical space and it is essential to ensure an extent of success in *hit* identification. However, despite its finity, chemical space is so vast (10^{26} synthesizable molecules) [97] that intense effort is necessary for charting it to explore the boundaries. This led to the development of concepts such as DOS (diversity-oriented synthesis) [98] and BIOS (biology-oriented synthesis) [99]. In screening campaigns of diverse libraries, the number of unique core structures is higher and therefore more informative about binding affinities. By consequence, diversification in library design greatly increases the probability of finding biologically active compounds. [100]

6.1.2 Chemical space and molecular properties

Chemical diversity can be characterized by structural diversity, side-chain diversity, functional group diversity, and stereochemical diversity. Such properties can be expressed by drug- or lead-likeness parameters as well as topological or pharmacophoric descriptors, which have been grouped into rules over time. Drug-likeness is mostly related to physico-chemical properties that impact

pharmacokinetics and bioavailability. They have e.g. an effect on the permeation through the cell membrane, so the considered properties describe the hydrophilicity/hydrophobicity balance and the molecular weight among other factors. The most widely used example of such rules is the Lipinski Rule of five (RO5) for orally administered drugs, [101] consisting of four major constraints:

- molecular weight ≤ 500
- Log P, octanol/water partition coefficient, ≤ 5
- H-bond donors, atoms bound to a hydrogen, ≤ 5
- H-bond acceptors, atoms without formal positive charge, ≤ 10

Molecules that stay within those limits are more prone to pass through the lipid bilayer of the cell membrane. [102] Thus, the absorption at the gastrointestinal level is increased and the oral administration effective. However, despite constituting a good starting point, the RO5 does not account for all important criteria in differentiating between drug- and non-drug-like, such as the number of rotatable bonds or the impact of selective transporters on the cell membrane. Additionally, it is worth mentioning that nowadays nearly half of the FDA-approved drugs are not compliant with the RO5. [103] Therefore, various adjustments followed the RO5. First of all, the Veber's rule includes two additional parameters: the number of rotatable bonds, excluding bonds to hydrogens, and the topological polar surface area (TPSA), calculated as the sum of fragments' contributions to the polarity of the molecule. According to the Veber's rule, the number of rotatable bonds and the TPSA should not exceed 10 and 140 \AA^2 , respectively. [104] Figure 20 shows the distribution of the mentioned descriptors in the Enamine database of 646 approved FDA drugs. [105]. The box plot presents the descriptors in the x-axis and the respective values on the y-axis. In Table 10, the statistical measures are summarized for each descriptor for clarity and it is notable that FDA approved drugs do not necessarily respect the RO5, as they show, for example, molecular weight up to 900 Da (instead of 500 Da) and number of hydrogen bonds acceptors up to 21 (instead of 10). In recent years, research in the field of natural products and macrocycles has gradually gained momentum. Despite such molecules largely exceeding the RO5 and other related drug-likeness rules, they can be actively transported inside the cell, so the RO5 is being gradually abandoned as a strict set of guidelines. [102]

Building blocks selection

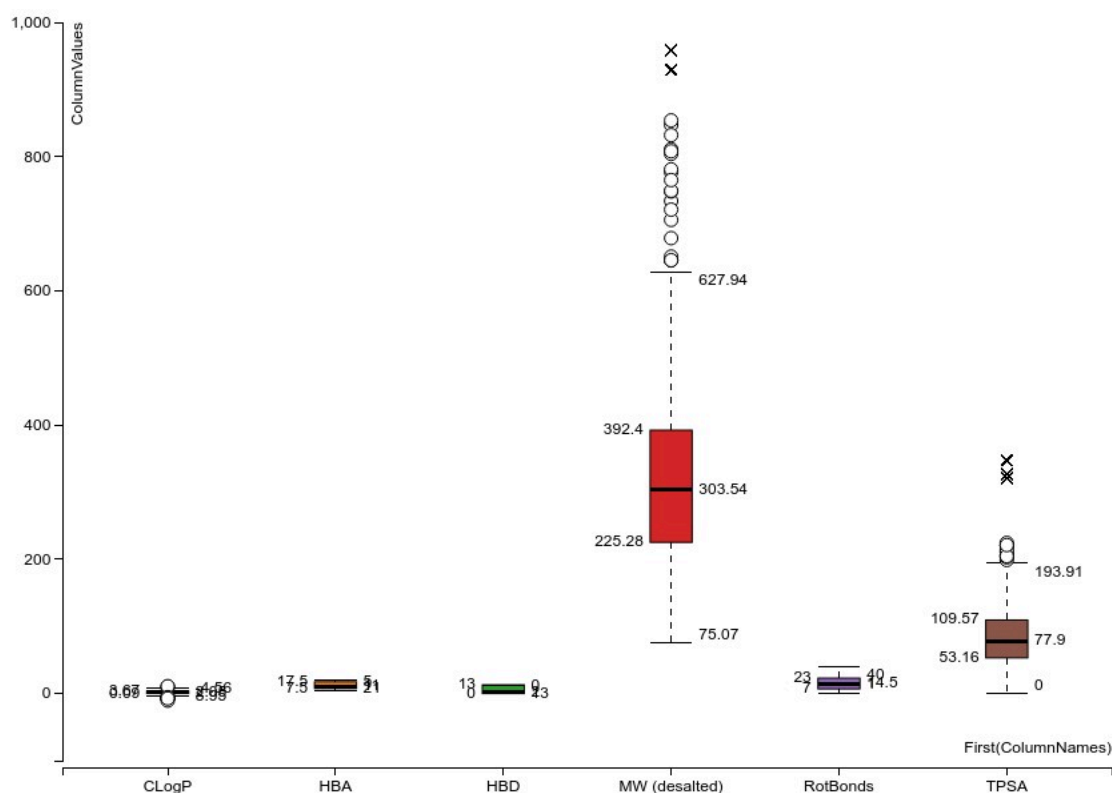


Figure 20: Box plot of the six descriptors included in the Lipinski's and Veber's rules for the Enamine database of FDA approved drugs. The descriptors are on the x-axis and the respective values on the y-axis. The values inside the box plot represent the median, the interquartile ranges and the whiskers (see Table 10).

Table 10. Statistical measures of the six descriptors included in the Lipinski's and Veber's rules, calculated for the Enamine database of FDA approved drugs.

Statistic measure	LogP	HBA	HBD	MW(desalted)	Rotatable bonds	TPSA
Minimum	-10.498	5.0	0.0	75.067	1.0	0.0
Lower Whisker	-4.56	5.0	0.0	75.067	1.0	0.0
Lower Quartile	0.085	7.5	0.0	225.284	7.0	53.16
Median	2.053	11.0	2.0	303.538	14.5	77.9
Upper Quartile	3.669	17.5	13.0	392.4	23.0	109.57
Upper Whisker	8.945	21.0	13.0	627.94	40.0	193.91
Maximum	10.967	21.0	13.0	958.224	40.0	347.32

Lower Quartile: value of the variables at 25% of the dataset; **Median:** value of the variables at 50% of the dataset; **Upper Quartile:** value of the variables at the 75% of the dataset; **IQR:** interquartile range, difference between the upper quartile and the lower quartile; **Lower Whisker:** 1.5 below the IQR; **Upper Whisker:** 1.5 above the IQR; **Minimum:** actual smaller value of the data set; **Maximum:** actual larger value of the data set.

Beside the drug-likeness related properties, other descriptors have been explored, which focus more on molecular substructures. Examples of such properties are the number of rings or heteroatoms in rings included in the cheminformatics package RDKit. [50] Those features are covered by the MQN (molecular quantum number) descriptors.[106] This set of features accounts for all significant structural properties related to atoms or connectivities and they were largely employed in medicinal chemistry to map the chemical space in search of novel drug-like compounds. [96] Lastly, molecules can be described by their three-dimensional shape, which is an important criterion related to binding affinity. This topic is discussed in detail in the section *Chemical space visualization*.

The diversity within the library members' space is connected with the reactants space due to the combinatorial connotation of DELs. [107]. Thus, to ensure heterogeneous collections of products, the building blocks must be carefully selected. This endeavour is especially challenging when huge data sets of fragments are provided by companies such as Enamine [36], Merck [37] or UCSF [108]. [109] Such databases contain millions of compounds and all are potentially valid building blocks for library synthesis. Therefore, cheminformatics tools come into the field for applying filters and selecting diverse sets of fragments.

6.1.3 Building blocks selection by cheminformatic tools

To date, the most noteworthy algorithm to design DELs is the *eDESIGNER*, developed by Martín, Nicolau and Toledo at Ely Lilly in 2020 [110]. The *eDESIGNER* focuses on the choice of the proper reactions to optimally connect building blocks defined by their functional group. Less attention is put on the selection of building blocks for diversity purposes and on the filters for big databases, beside special functionalities that react with water and would cause problems in a DEL context. In a recent publication from Prof. Varnek's group [111] which attempted to characterize the potential DEL chemical space, another algorithm called *Synthl* [112] was employed prior to the *eDESIGNER* for selecting the building blocks. These two publications show the importance of the number of functionalities in each building block. In particular, bifunctional or trifunctional fragments are preferred in the first steps (cycles) of the DEL synthesis, to enable successive library synthesis. On the other hand, bifunctional building blocks are necessary in the first cycle to be coupled to the DNA tag, whereas abundantly available monofunctional building blocks can only be used in the last cycle of a DEL synthesis, or if trifunctionalized scaffolds had been coupled in the first step.

In designing a DEL, especially if the diversity is the main priority, special attention needs to be paid to the scaffold, as it has been proved that scaffold diversity impacts the effectiveness of the library more than the building blocks diversity. [15],[113] Therefore, the filters applied to large collections of building blocks for selection are based on the chemistry that characterize the library. However, few criteria can be considered quite general and widely applicable. Fragments displaying functionalities that can disturb biological assays, called PAINS (Pan-assay interference compounds) are to be avoided. [114] The second group of criteria is embedded in a rule, similar to the RO5, but adapted to fragments, the Goldberg's rule of two, which quotes: MW < 200 Da, cLogP < 2, number of H-bond donors ≤ 2, and number of H-bond acceptors ≤ 4. Following this rule may ensure an acceptable compound size and compliance with the RO5 of the final library members. [115] Once the building blocks are adapted to the chemistry and they conform with the Goldberg's rule of two, they can be selected for diversity.

6.1.4 Chemical space visualization and similarity

Diversity is by definition the opposite of similarity. The more intuitive way to understand the similarity between molecules is to project them in the chemical space according to the values of the descriptors and calculate the Euclidean distance as presented in Equation 2. Here, D is the distance or the plane in a three-dimensional space, and a and b the projection of the molecules on the axes of the chemical space. [116]

$$D = \sqrt{a^2 + b^2} \quad (\text{Eq. 2})$$

6.1.4.1 Principal component analysis (PCA)

The chemical space is delimited by chemical properties that represent its dimensions. Therefore, for visualization and interpretation, up to three variables can be easily employed, whereas more than that becomes complicated. The MQN (molecular quantum number) descriptors constitute a good example for this concept as they account for 42 variables which are all equally important for describing molecules. In such cases, reducing the dimensions of the data is essential and the most practised algorithm for this purpose is the PCA (principal component analysis). [117] The PCA identifies the variables that impact the most the considered dataset and, by linearly combining them, it generates new variables called *principal components* or *eigenvector*. In particular, the first component coincides with the variable with the highest variance, the second component is orthogonal to the first and the rest likewise.

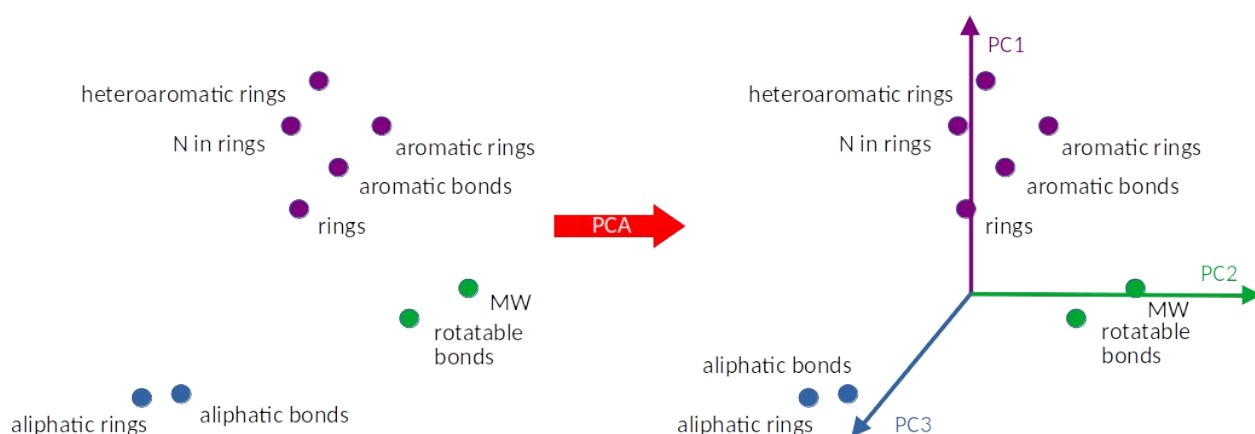


Figure 21: PCA. The variables describing a data set (nine dimensions) are linearly combined into the principal components or eigenvectors. The dots in the 3D space represent the variables and they are colored with the same hue if correlated. After the PCA, the initial variables are grouped according to correlation and form the new principal components.

The dots in Figure 21 are variables which describe structural features of a set of molecules. They are placed in the 3D space, defined by three orthogonal planes, according to their values and correlated variables are depicted in the same hue. Some correlations are clear: for example the variables describing aromatic rings and aromatic bonds (bonds in aromatic rings) show similar information, as well as the two variables rotatable bonds and molecular weight. In this model, there are nine variables and their values cannot be plotted together as nine dimensions are too many. Therefore, the PCA comes into play and generates three principal components which account for all the nine variables. In this way, it is possible to display a dataset with many dimensions. In Figure 22, the dots in the chemical space defined by the three principal components are molecules projected in the chemical space according to the values of the variables combined in the principal components. Peculiar characteristics can be deduced by looking at the plot in Figure 22. For example, molecules on the bottom left in Figure 22 are characterized by aliphatic substructures while molecules on the top feature aromatic substructures. With this procedure, complex data sets can be plotted aiding data analysis.

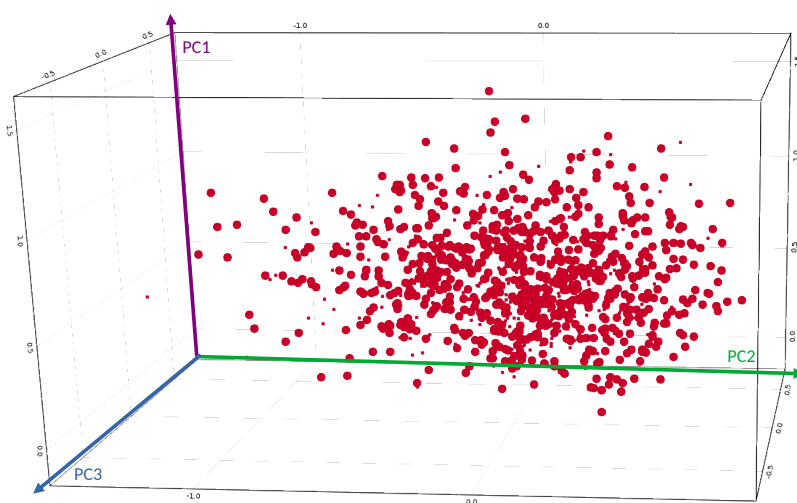


Figure 22: The chemical space is delimited by the three *principal components* extracted by PCA. In this scatter plot, the data points represent the Enamine database of 646 FDA approved drugs, characterized by the properties included in the principal components.

6.1.4.2 Principal moment of inertia (PMI)

Beside chemical properties, molecules can be described by geometry as well. This parameter is as important as the chemical descriptors because molecules bind to target proteins when their shape is complementary to the binding pocket. [56] Although basing the binding affinity only on shape would be an oversimplification, this parameter can be considered to obtain a more comprehensive overview on chemical space coverage by a screening library. The shape of the molecules can be described as a combination of the three principal moments of inertia (PMIs), which correlate the shape and size of the molecules with the axes of inertia. In Figure 23 A, as an example, the PMIs of the water molecule are depicted in relation to the centre of mass of the molecule. The centre of mass of the molecule is defined as the point in which the sum of all the distributed masses equals to zero. [56] The normalization of the first two PMIs over the third one generates the triangle plot displayed in Figure 23 B. In this case, the considered dataset is the Enamine database for FDA approved drugs. At the vertices of the triangle are the three extreme shapes (rod, sphere and disc) and the molecules can assume all shades of those three. It is noticeable that, with some exceptions, approved drugs account predominantly for the rod-like shape and slightly for the disc-like.

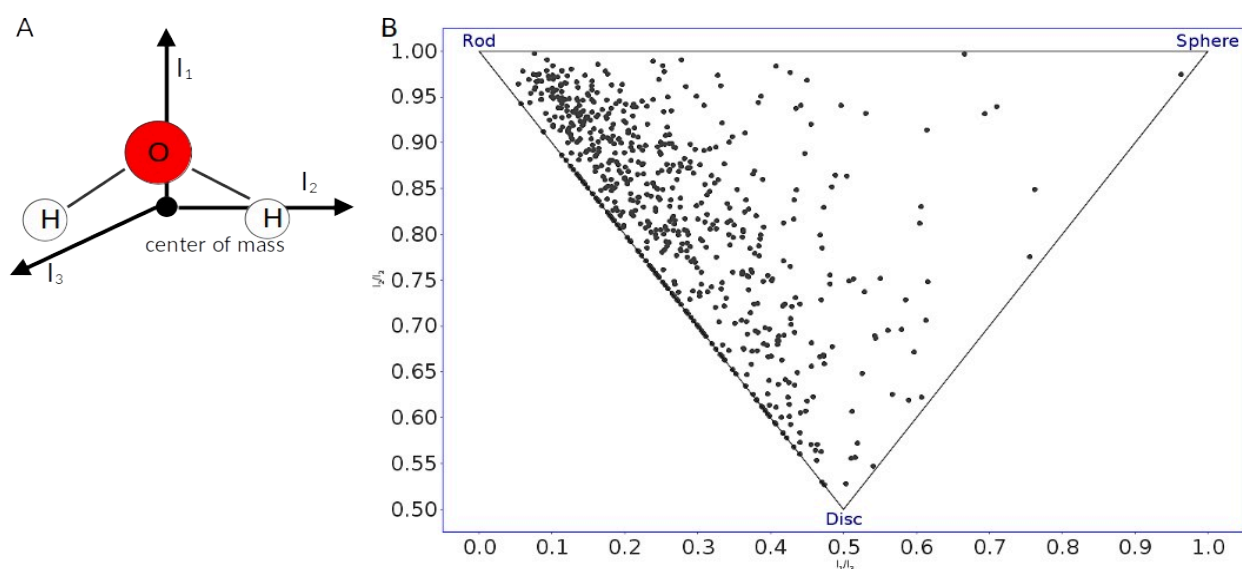


Figure 23: PMI analysis. (A) PMIs for the water molecules, as example, in function of the centre of mass. The origin of the three PMI axes (I_1 , I_2 and I_3) is the centre of mass. (B) The triangle plot delimited by the normalized values of the first two PMIs over the third, where the molecules are placed according to their shape: rod-, disc- or sphere-like. The data points are the Enamine database of FDA approved drugs.

In the context of combinatorial libraries, such as DELs, and especially for untargeted libraries, diversity of the members is an essential feature to enhance the probability in identifying relevant *hits*. This can be achieved by accurately selecting the building blocks prior to library synthesis from larger collections of fragments. To date, no selection tool which aim for diversity has been reported. Additionally, visualization options that assess the significance of the virtual library prior to synthesis could serve in optimizing resources and efforts. In fact, by observing the chemical space covered by the virtually designed library, it could be possible to improve decision making and to prioritize chemistry and building blocks before starting the synthesis.

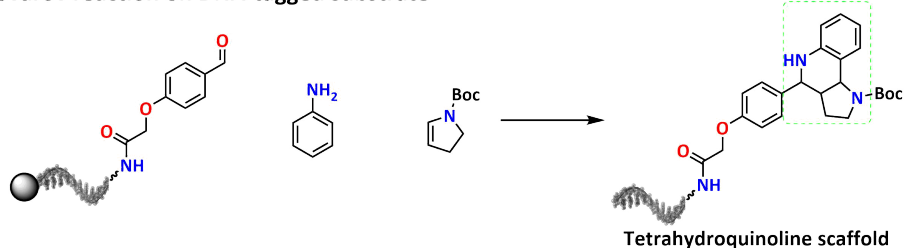
6.2 Aim

The objective of this chapter is to develop an algorithm which automatizes the selection of building blocks for DNA-encoded libraries. It should contain all the filters reflecting the above mentioned rules such as the Goldeberg's rule of two, to reduce the manual effort for building block selection to a minimum. However, the chemist's intervention is needed to input the chemistry which connects the building blocks. In fact, the building blocks selection is based on the chemistry characterizing the library and on the cycle for which the building blocks are selected. As mentioned before, the number of functionalities depends on the cycle. After connecting the building blocks according to the designed chemistry and forming the members of the library, the workflow should present the data in graphs which enhance the interpretation of the results and guide further decisions. Aiming for diversity, the library members should cover a large chemical space. This translates, for example, into occupying large portions of a scatter plot defined by the PCA (principal component analysis) dimensions obtained combining molecular properties. Additionally, the library members should vary in shape and geometry and this parameter can be verified by the PMI (principal moment of inertia) analysis. In order to assess their significance, virtually generated libraries can be compared to drug-like molecules in terms of chemical space and shape and, to render the whole workflow accessible to chemists, the open source KNIME Analytics Platform is the software of choice.

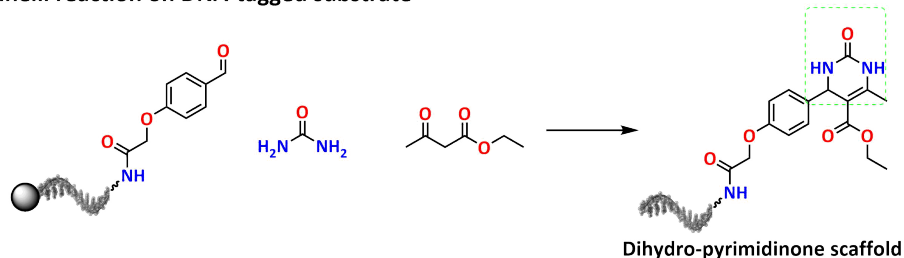
6.3 Methods

In our laboratory, three reactions were tested by Dr. Marco Potowski (postdoctoral fellow in the PD Brunschweiler's group) on DNA conjugates, namely the Povarov, Biginelli and *aza*-Diels-Alder reaction (Scheme 4). [22]

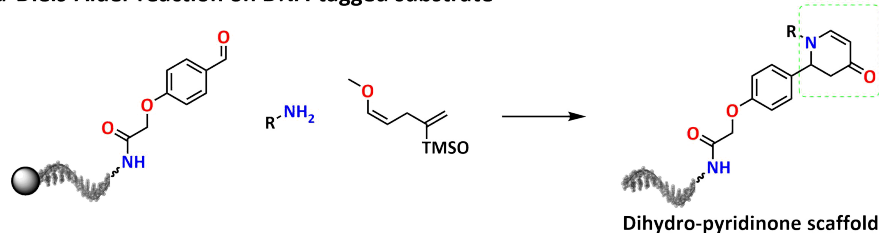
Povarov reaction on DNA-tagged substrate



Biginelli reaction on DNA-tagged substrate



aza-Diels-Alder reaction on DNA-tagged substrate



Scheme 4. Reactions tested on DNA-tagged substrates: Povarov, Biginelli and *aza*-Diels-Alder. In the dotted square the respective scaffolds are emphasized.

Due to the good yields and wide substrate scope, they were considered as potential candidates for library design and synthesis. For the purpose of creating three virtual libraries based on the aforementioned reactions, the building blocks were selected from the Aldrich Market Select database via the KNIME Analytic Platform. This database was chosen because it contains a comprehensive list of commercially available building blocks. In KNIME, the molecular weight threshold was set to 200 Dalton following the rule of two. Building blocks with secondary functional groups, which could yield

side products were excluded. Another regarded criterion in the selection was the substitution pattern on the aromatic ring, wherein sterically bulky substituents and electron withdrawing groups in the *ortho* position were ruled out. The filters were also based on the reactivity of some specific building blocks such as heteroaromatic amines which were therefore removed from the data set. After selecting all the necessary classes of molecules to build the virtual libraries, the chemistry-oriented extensions of KNIME were used for simulating the chemical reactions. Although experimentally multi-component reactions occur in one pot, they were split in series of subsequent steps to obtain the final molecules. The size of the three libraries was kept more or less constant to allow for comparison at later stages. The properties of each library member were then determined in order to predict the final chemical and geometrical space coverage. We choose to define the chemical space by physicochemical features grouped into 42 topological and pharmacophoric descriptors called MQN because we considered them to be very comprehensive. [106] The complete list of MQN descriptors with the respective meaning in terms of count of atoms or substructures is explicated in Table 11.

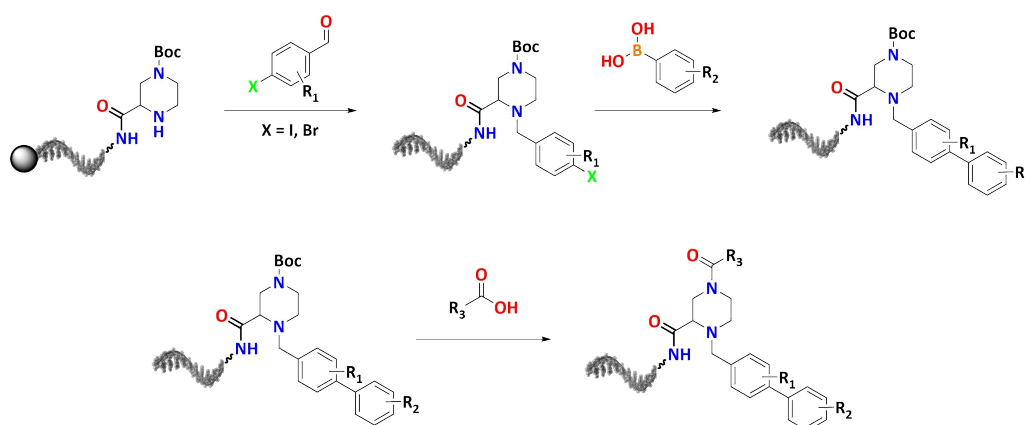
Table 11: Overview on the MQN descriptors and the respective meaning in terms of count of atoms or substructures.

MQN descriptors	counts	MQN descriptors	counts
MQN1	carbon	MQN22	H-bond donor sites
MQN2	fluorine	MQN23	H-bond donor atoms
MQN3	chlorine	MQN24	negative charges
MQN4	bromine	MQN25	positive charges
MQN5	iodine	MQN26	acyclic single valent nodes
MQN6	sulfur	MQN27	acyclic divalent nodes
MQN7	phosphorous	MQN28	acyclic trivalent nodes
MQN8	acyclic nitrogen	MQN29	acyclic tetravalent nodes
MQN9	cyclic nitrogen	MQN30	cyclic divalent nodes
MQN10	acyclic oxygen	MQN31	cyclic trivalent nodes
MQN11	cyclic oxygen	MQN32	cyclic tetravalent nodes
MQN12	heavy atoms	MQN33	3-membered rings
MQN13	acyclic single bonds	MQN34	4-membered rings
MQN14	acyclic double bonds	MQN35	5-membered rings
MQN15	acyclic triple bonds	MQN36	6-membered rings
MQN16	cyclic single bonds	MQN37	7-membered rings
MQN17	cyclic double bonds	MQN38	8-membered rings
MQN18	cyclic triple bonds	MQN39	9-membered rings
MQN19	rotatable bonds	MQN40	≥10-membered rings
MQN20	H-bond acceptor sites	MQN41	nodes shared by ≥2 rings
MQN21	H-bond acceptor atoms	MQN42	edges shared by ≥2 rings

The geometrical space was defined by the PMI (principal moment of inertia) properties which are related to the three-dimensional shape of the molecules. This prediction could function as feedback for the building blocks selection as it could underline the effect of the starting materials on the properties of the final molecules. The properties of the final library members were compared with

Enamine catalogues of drug-like compounds since it provides an estimate for the value of the chemical and geometrical space coverage of the potential library. The comparison may lead to two possible scenarios: on the one hand, the library chemical and geometrical space might overlap with existing databases, meaning that the molecules produced by the library might show similar properties and presumably similar biological activity. This outcome might be considered promising in terms of *hit* identification rate but does not produce novelty. If, on the other hand, the two data sets cover different chemical and geometrical space, this characteristic might pave the way for discovering new classes of bioactive molecules or for expanding the biological space by exploring not-yet drugged targets. Furthermore, such comparisons allow for better allotting resources in the laboratory practice. One could prioritize the synthesis of library covering either a wider or a very specific chemical and geometrical space according to the current requirements.

A very similar procedure was applied to the creation of another virtual library based on subsequent reductive amination and Suzuki coupling (Scheme 5).



Scheme 5. Design of the library based on reductive amination and Suzuki coupling. The initial building block is the diamine which is coupled to the DNA via amide bond. The aldehyde with the halide moiety reacts by reductive amination with the free amine and the Suzuki coupling takes place with the boronic acid. The last step is the final coupling with a carboxylic acid after *boc*-deprotection of the second amine.

The starting materials, namely arylaldehydes containing a halide moiety, were selected from the Enamine REAL space database and the molecular weight threshold of 200 Da was applied as well. For the purpose of optimizing the building blocks selection and predicting the properties of the final molecules, the library was enumerated according to the selected starting materials and the resulting coverage of chemical and geometrical space was compared with three Enamine collections: the FDA approved drugs, natural products- and lead-like compounds. For this second application of the

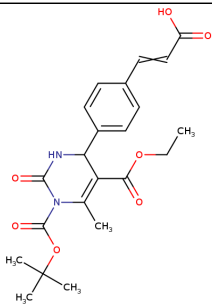
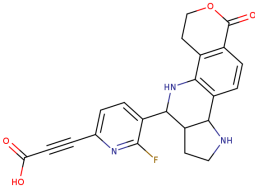
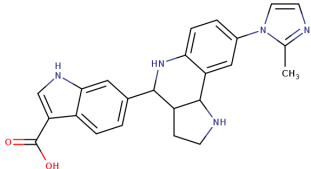
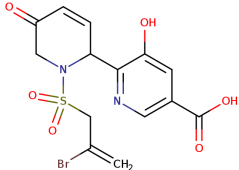
enumeration procedure, the descriptors were chosen to be closely related to the Lipinski's RO5 and the Veber's rule: the molecular weight (MW), the hybrid LogP (SlogP), the number of hydrogen bond donors and acceptors (HBD and HBA), the topological polar surface area (TPSA) [118] and the number of rotatable bonds. The SlogP (hybrid logP) was chosen because, in addition to the conventional atomic LogP, it takes into consideration the effect of the neighbouring atoms and is therefore more suitable for more complex systems. [119] With only six variables to delimit the chemical space, the dimensionality reduction by PCA (principal component analysis) was not essential and was therefore omitted. This omission was possible because the descriptors could be alternatively employed as axes in the scatter plot defining the chemical space and thus each molecule in the plot could be directly correlated to the value of the considered descriptor. Furthermore, the usage of the PCA causes a loss of information and interpretability, since the PCA dimensions hardly reflect a clear combination of the initial variables. Therefore, for this study case, two maps of the chemical space were generated by plotting the MW vs HBD vs HBA and the SlogP vs TPSA vs rotatable bonds, respectively. Comparing the chemical space covered by those data sets was intrinsically informative as we expected natural product-like molecules to cover a different space than drug-like compounds. Via this similarity assessment, it was possible to clarify whether the library members were in the right direction to become drugs after an optimization process. Moreover, the coverage of similar chemical space might hint to a similar bioactivity and this information might be beneficial in choosing the target proteins to screen during the selection assay. If the virtual library members and the ligands for a specific protein resemble in terms of chemical space, this protein might constitute an attractive option for screening.

6.4 Results and discussion

6.4.1 Libraries based on Povarov, Biginelli and *aza*-Diels-Alder reactions

After appropriately selecting the building blocks, the three libraries based on tetrahydroquinoline (Povarov, P), dihydro-pyrimidinone (Biginelli, B) and dihydro-pyridinone (*aza*-Diels-Alder, DA) scaffolds were enumerated with KNIME (examples shown in Table 12). In Table XX, the molecules are depicted with the respective affiliations.

Table 12: Example molecules extracted from the P, B and DA libraries.

Molecule	Scaffold
	Biginelli
	Povarov
	Povarov
	<i>aza</i> -Diels-Alder

The MQN descriptors were calculated for randomly picked samples of 1000 molecules for each library and for the Enamine drug-like compounds database and the dimensionality was reduced from 42 to three by PCA. The scatter plot in Figure 24 illustrates the chemical space defined by the PCA dimensions resulting from the dimensionality reduction and normalization of the four data sets. In particular, the x-, y- and z- axes represented the PCA dimension 0, 1 and 2 respectively, namely the three *principal components*. Overall, the four databases overlap to some extent, especially in the centre of the plot. The subset that overlaid the most with the drug-like molecules space was the DA occupying in prevalence the centre, whereas the Biginelli and Povarov reactions-based libraries covered the bottom left and right areas of the plot, respectively. The Biginelli library settled a lower value of PCA dimensions 0 and 1, while the Povarov library covered lower values of the PCA dimension 1 and higher values of the PCA dimension 0. The Enamine data set stood out for occupying regions with higher values of the PCA dimension 1.

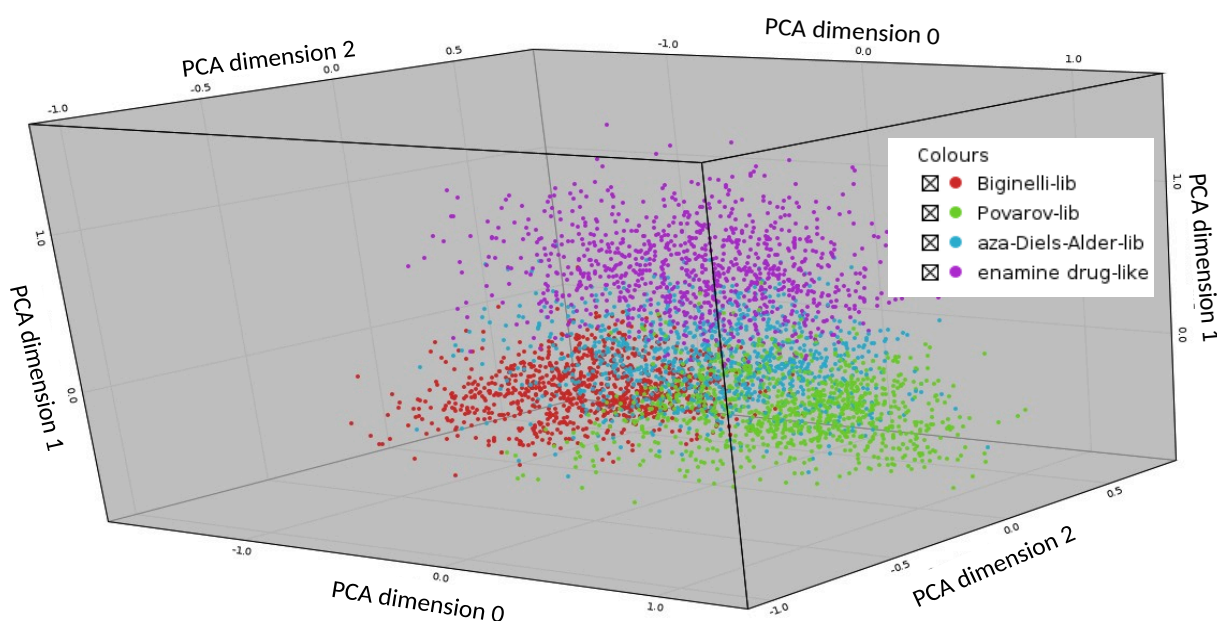


Figure 24: Scatter plot of the distribution of the P, B and DA libraries in comparison with the Enamine database of drug-like compounds. The dimensions of the scatter plot are represented by the principal components from the PCA.

To understand in depth the meaning of this projection, the PCA dimensions needed to be tracked back to the initial variables that influenced them the most. After extracting the three most influential *eigenvector*, corresponding to the three PCA dimensions, the variables connected to them were analysed. The correlations between the eigenvectors and the three most influencing variables each are

depicted in Table 13 and the analysis resulted in the following assumptions. The eigenvector 0 was related mostly to acyclic single bonds, cyclic divalent nodes, acyclic single valent nodes, whereas the eigenvector 1 to heavy atoms, cyclic trivalent nodes, hydrogen bond acceptor atoms and eigenvector 2 with acyclic triple bonds, 5-membered rings, cyclic nitrogen. By consequence, the PCA dimensions were considered to be affected by the same parameters. To give a more chemical interpretation, the first principal component seemed to prioritize aliphatic structures as well as saturated rings. The second principal component could relate to the presence of heavy atoms which acted as HBA (hydrogen bond acceptor), such as carbonyl oxygen. Notably, heavy atoms in this context represent all atoms except hydrogen. Finally, the third principal component could be correlated with nitrogen-containing heterocycles, in particular 5-membered rings.

Table 13: Correlation of each principal component (or eigenvector) with three initial variable, in order to interpret the PCA.

eigenvector	MQN descriptor
0. eigenvector	acyclic single bonds
0. eigenvector	cyclic divalent nodes
0. eigenvector	acyclic single valent nodes
1. eigenvector	heavy atoms
1. eigenvector	cyclic trivalent nodes
1. eigenvector	H-bond acceptor atoms
2. eigenvector	acyclic triple bonds
2. eigenvector	5-membered rings
2. eigenvector	cyclic nitrogen

With this in mind, the position of the Enamine data set on the upper part of the graph meant that drug-like molecules contained a higher number of heavy atoms in comparison to the library members. On the other hand, the comparison between the Povarov and Biginelli libraries confirmed the nature of the two different scaffolds. In fact, the former contained preferentially saturated rings compared to the latter. Overall, in terms of structural properties, the *aza*-Diels-Alder library placed itself the closest

to the drug-like space, representing an excellent candidate to be investigated further with different combinations of building blocks. This finding was also confirmed in terms of geometrical properties as shown in the PMI plot in Figure 25.

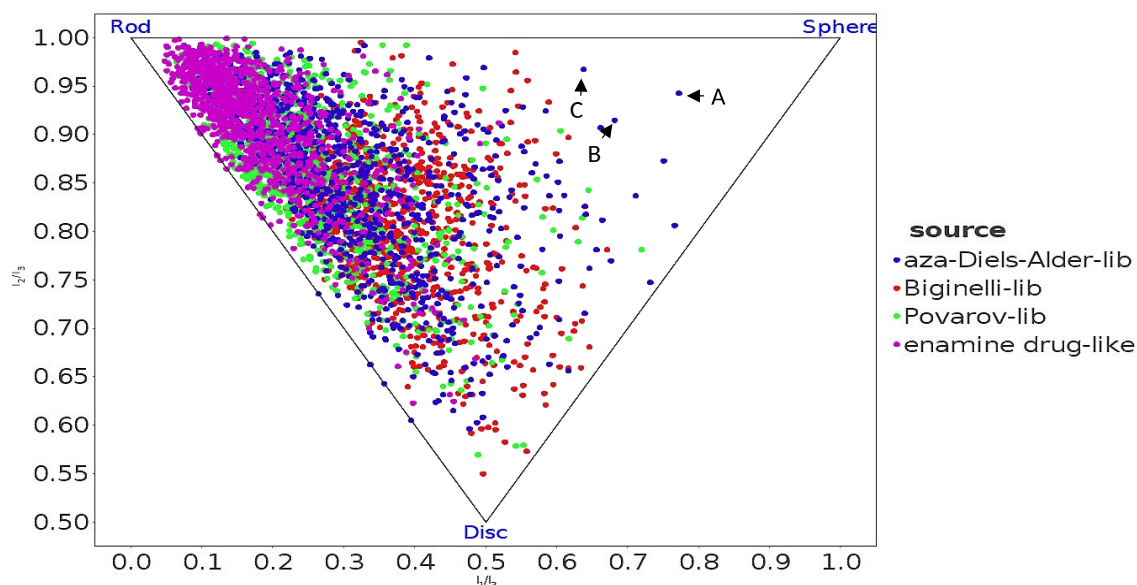


Figure 25: PMI triangle plot for the three P, B and DA libraries and the Enamine database of drug-like compounds. Particular molecules are highlighted by the letters A, B and C (structures shown in Figure 26).

Here, the dots were colored according to the affiliation of the molecule to the database: the three libraries or the Enamine database of drug-like compounds. It is noticeable that drug-like compounds tend to take the shape of a rod and this common inclination was also detectable for the other data sets. Despite that, the *aza-Diels-Alder* library, containing the highest Csp^3 fraction, presented some examples of spherical compounds (Figure 26), followed by the Biginelli and Povarov libraries. The roundness of the molecule depends on the occupancy of the 3D space in all directions. Molecules with aromatic rings tend to be more two-dimensional since their atoms are disposed in planes, whereas molecules with sp^3 bonds stretch over all directions of the space.

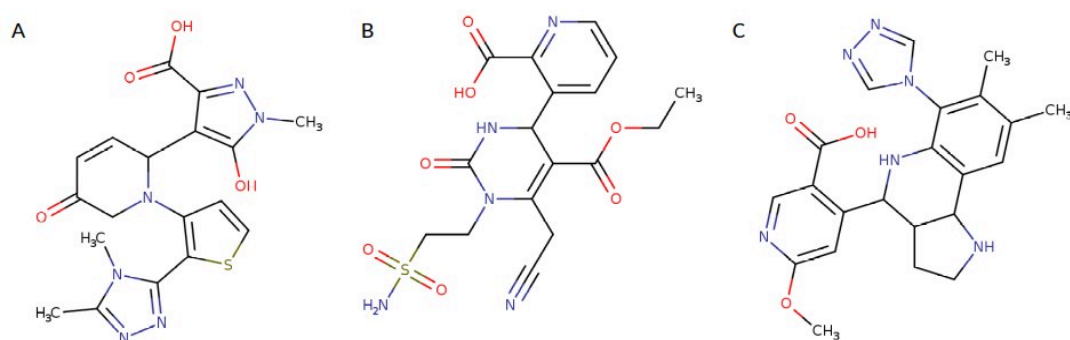
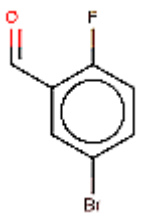
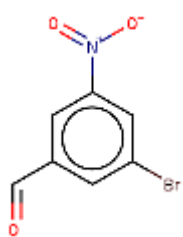
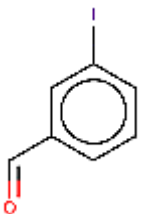
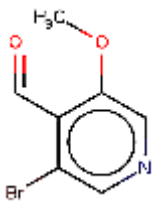
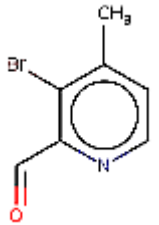


Figure 26: Molecules corresponding to the data points in the PMI plot in Figure XXX. (A) DA molecule with the most spherical shape, followed by (B) the B molecules and (C) the P molecule.

6.4.2 Library based on reductive amination and Suzuki coupling

The building blocks, in this case, were extracted from the Enamine REAL database and filtered according to different criteria. First of all, the required functional groups to react in the library were the aldehyde and the Bromine/Iodine groups for the reductive amination and the Suzuki coupling, respectively. From this set, molecules containing sulphur atoms were excluded as well as PAINS and fragments including more than one ring as they would be too heavy for one cycle in the library synthesis. After that, among additional filtering steps, we ensured that no sterically hindering moiety was placed in *ortho* position to the aldehyde and we picked a diverse set of five as shown in Table 14 where the related properties are displayed as well. As noticeable, the building blocks were diversely substituted and heteroaromatic rings were present to increase the diversity of the library. The Bromine or Iodine substituents were placed in different positions on the aromatic ring in order to expand the shape scope, as they could orient the ring coupled by the Suzuki reaction in different directions.

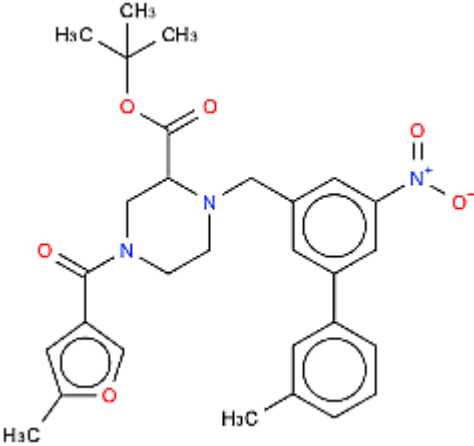
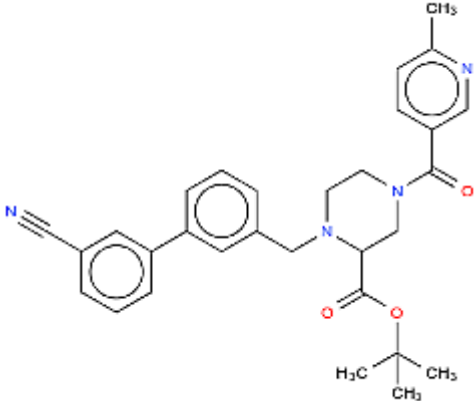
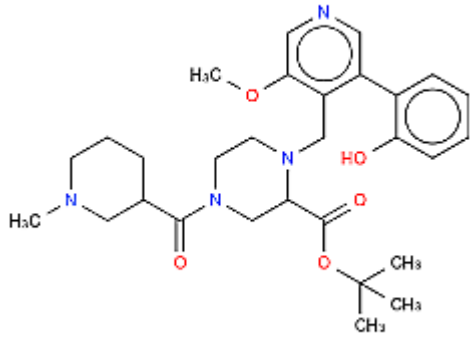
Table 14: Five selected aldehydes with halogen moieties, with few respective properties.

Molecule	Catalogue_ID	ExactMW	NumRings	NumAromatic Rings	NumAromatic Heterocycles
	EN300-20799	201.94	1	1	0
	EN300-305394	228.94	1	1	0
	EN300-112542	231.94	1	1	0
	EN300-300047	214.96	1	1	1
	EN300-345166	198.96	1	1	1

These building blocks were virtually reacted with a mono-*boc*-protected diamine by reductive amination and then with randomly selected boronic acids from the same database via Suzuki coupling. The last step of the virtual synthesis was a *t-boc*-deprotection/amide coupling virtual reaction in which

the *t-boc* (tert-butyloxycarbonyl) group on the *boc*-protected amine was substituted by the carboxylic acid. The final products were properly labelled according to the utilized building blocks to ensure traceability over the whole library synthesis and screening process. In Table 15, a random sample of 5 molecules is illustrated with the respective labels. The molecules are composed by four parts which derive, respectively, from the aldehyde, the diamine, the boronic acid and the final carboxylic acid. Notably, on the diamine side, there is another *t-boc* group, which correspond to the attachment point to the DNA. The "*BB_comb*" column contains the labels which are characterized by three alphabet letters A, B and C for cycle 1, 2 and 3 respectively and by numbers defining the specific substructure. This labelling step is essential with DELs, and combinatorial libraries in general, due to the vast numbers of compounds in the data set. The same labelling procedure could be applied to a 96-well plate in which each compound is labelled according to the respective well.

Table 15: Examples of molecules resulting from the reactions between the five aldehydes and boronic acids, with respective labels for traceability.

BB_comb	Molecules
A2_B3_C53	 <p>Chemical structure of molecule A2_B3_C53. It features a central piperazine ring. One nitrogen atom is substituted with a tert-butyl ester group (CH₃C(CH₃)₂COO-). The other nitrogen atom is substituted with a 4-(4-methylphenyl)phenyl group. The piperazine ring is also substituted with a 5-methylfuran-2-carbonyl group and a nitro group (NO₂).</p>
A3_B15_C22	 <p>Chemical structure of molecule A3_B15_C22. It features a central piperazine ring. One nitrogen atom is substituted with a tert-butyl ester group (CH₃C(CH₃)₂COO-). The other nitrogen atom is substituted with a 4-(4-cyanophenyl)phenyl group. The piperazine ring is also substituted with a 4-methylbenzamide group and a 4-methylpyridin-2-ylmethyl group.</p>
A4_B10_C73	 <p>Chemical structure of molecule A4_B10_C73. It features a central piperazine ring. One nitrogen atom is substituted with a methyl group (H₃C). The other nitrogen atom is substituted with a 4-(4-methylphenyl)phenyl group. The piperazine ring is also substituted with a 4-methylpyridin-2-ylmethyl group, a hydroxyl group (HO), and a tert-butyl ester group (CH₃C(CH₃)₂COO-).</p>

The descriptors for those molecules were calculated and, in contrast to the previous analysis, only descriptors related to the Lipinski RO5 and the Veber's rule were considered such as the number of HBD (hydrogen bond donors), the number of HBA (hydrogen bond acceptors), the SlogP (hybrid LogP), MW (molecular weight) and TPSA (topological polar surface area). In order to improve the interpretability of the results, the PCA (principal component analysis) was omitted and the descriptors were alternatively plotted in the 3D chemical space. The synthesized and described library was compared to the Enamine collections as illustrated in Figure 27-30. In particular, the Enamine collections included FDA approved drugs as well as natural products-like and lead-like compounds. Lead-like compounds differ from drug-like compounds as they normally show lower molecular weight, which allow for further optimization in terms of pharmacokinetics. An overview of the six descriptors over the four data sets is illustrated in Figure 27. It is noticeable that drugs surpass all other descriptors from all points of view, whereas the library and the lead-like compounds are the most similar. They coincide regarding the TPSA and the number of rotatable bonds, while they are very close for MW, NumHBA and SLogP. This is certainly a good sign, as DEL *hits* are submitted to optimization steps that would, for example, increase the MW or the TPSA to increase affinity and bioavailability.

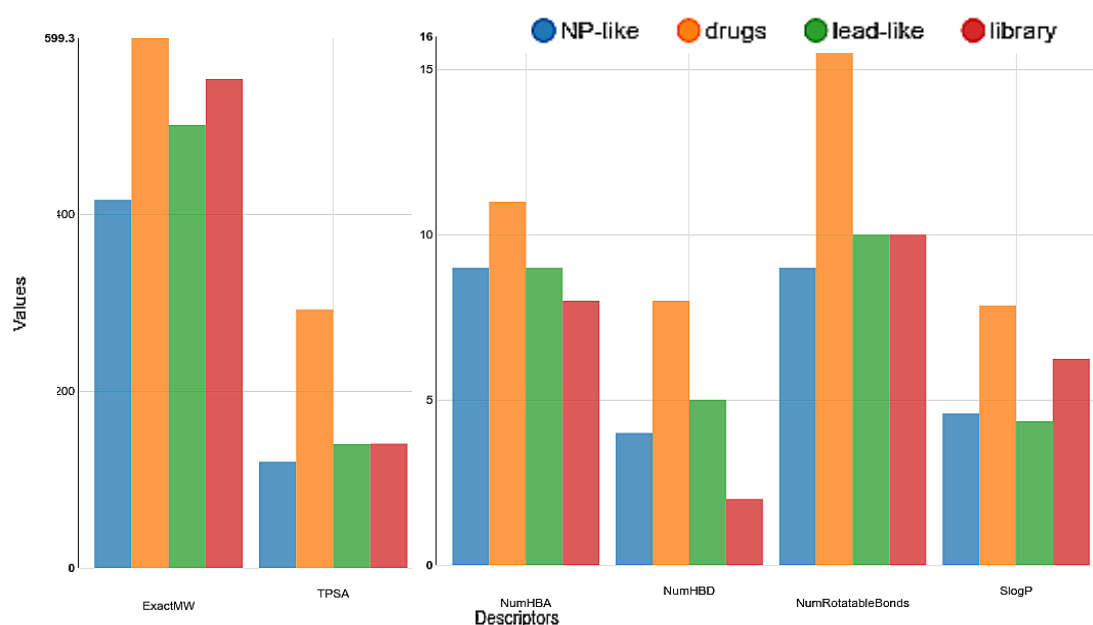


Figure 27: Descriptors values over the four data sets: library (red), Enamine databases of FDA approved drugs (orange), lead-like (green) and natural products-like (blue) compounds. On the x-axis the descriptors are listed, while on the y-axis the respective values are represented. The bar chart is split in two because the MW and the TPSA would be out of scale, rendering the differences among the other descriptors invisible. NP = natural products, MW= molecular weight, TPSA = topological polar surface area, NumHBA = number of hydrogen bond acceptor, NumHBD = number of hydrogen bond donors.

In 28 the four data sets are depicted in different colours in the 3D chemical space defined by the SLogP, MW and TPSA.

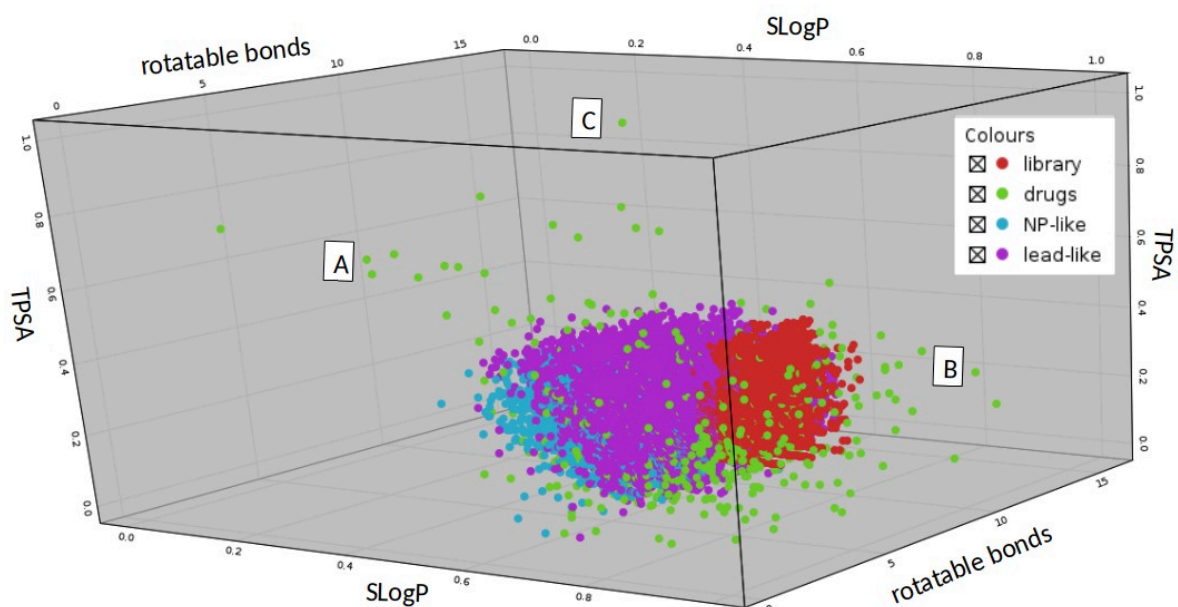


Figure 28: 3D scatter plot delimited by the variable SLogP, TPSA and number of rotatable bonds. Particular molecules are highlighted by the letters A, B and C (structures shown in Figure XXX). Library members are colored in red, FDA approved drugs in green, natural products-like (NP-like) and lead-like compounds in blue and purple respectively.

Regarding these three variables, the four data sets mostly overlap, except some molecules from the drug-like database that presented either a very hydrophilic structure, such as modified nucleotides (Figure 29 A), with negative SLogP or long saturated chains that made the number of rotatable bonds increase drastically (Figure 29B). The variable TPSA compressed the data points in one region of the chemical space, meaning that in terms of balance between polar and unpolar groups the four data sets were very similar. Only the drug-like molecules differed and scattered away from the densely populated area, reporting examples of very high TPSA. In fact, the exemplary molecule pictured in Figure 29 C features predominantly the polar oxygen and nitrogen atoms.

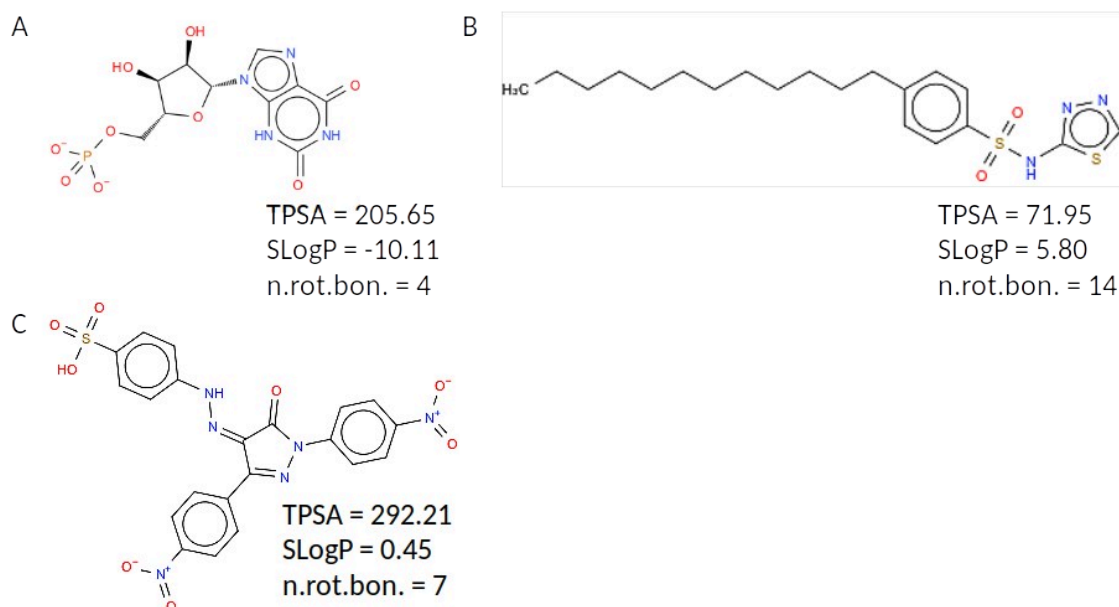


Figure 29: Molecules corresponding to the data point highlighted by the respective letters in the scatter plot in Figure 28: (A) compound with low SLogP, (B) compound with high number of rotatable bonds and (C) compound with high TPSA. These molecules are all examples of the Enamine database of FDA approved drugs. n.rot.bon. = number of rotatable bonds.

In Figure 30, the same scatter plot is displayed but, in this case, the axes represent the molecular weight, the numbers of hydrogen bond donors and acceptors. One feature that stands out was the presence of linear patterns due to the integer nature of the HBD and HBA descriptors, which can assume only whole numeric values. In contrast to Figure XX, here the three data sets are not overlaid and, while the drug database is spread all over the chemical space, the library members, the natural products- and the lead-like molecules are differentiated by the molecular weight. In fact, the library members cover higher molecular weights compared to the other data sets and average values of HBA. Notably, the drugs database stood out for providing HBA and HBD rich compounds. A positive outcome in this case is represented by the similarity between the library and the lead-like compounds, especially in the centre of the plot. This means that the library members are similar in molecular properties to lead-like compounds, which represent the penultimate stage in the drug discovery process.

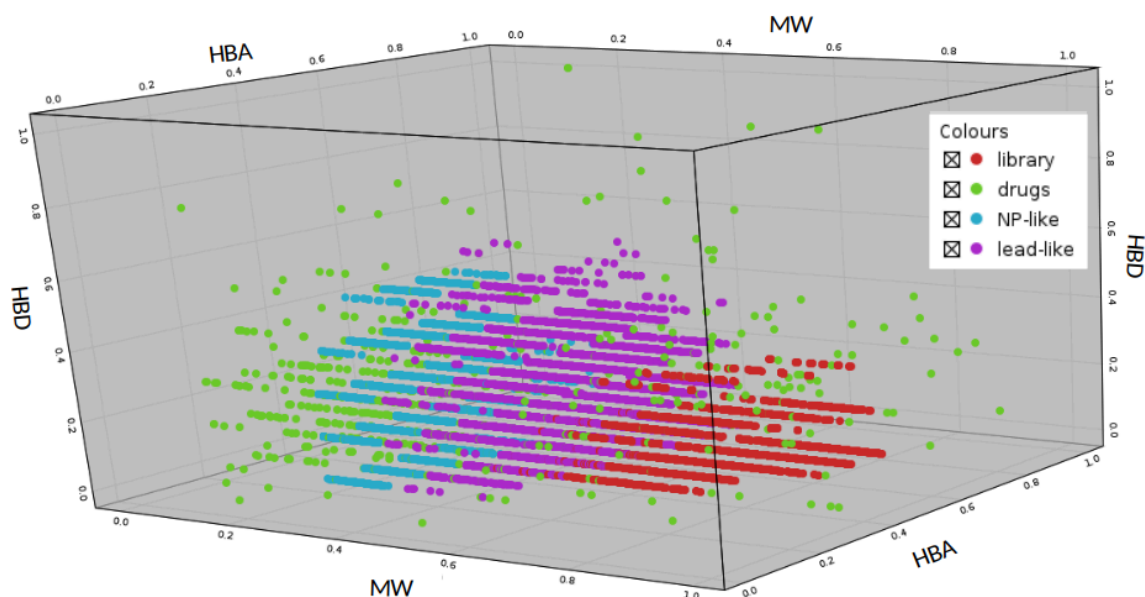


Figure 30: 3D scatter plot delimited by the variables MW, HBD and HBA and representing the Suzuki-based library in comparison with the Enamine databases of FDA approved drugs, lead- and natural products-like (NP-like) compounds.

It appeared clear from this second analysis that the results were more easily interpretable than after reducing the dimensions by PCA. In designing a not targeted library, the choice of building blocks that allow for wide chemical space coverage and for the similarity of library members with existing drugs or drug-like molecules are essential parameters to be considered.

Similar to the first case study, we analysed the libraries from a geometrical point of view according to the PMI properties (Figure 31). For this analysis the library members were compared to databases of existing drugs, lead-like compounds, and natural products-like molecules. The presence of natural product-like compounds at the edge between the rod- and disc-like shapes delineate structures showing linear structures, as depicted in Figure 32 A. Although the four data sets take mostly shapes between the rod and the disc, the library appeared to be the only set pointing towards the spherical shape (Figure 32 B). This feature is highly desirable in a library for expanding the scope of proteins that could interact with its members.

Building blocks selection

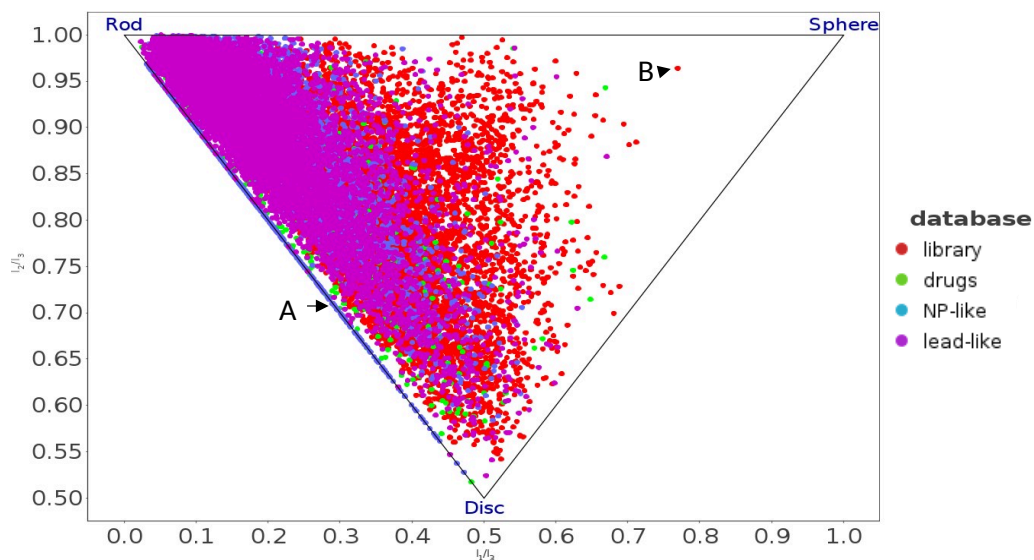


Figure 31: PMI triangle plot of the library base on reductive amination and Suzuki coupling compared with the Enamine databases of approved drugs, lead- and natural product-like compounds.

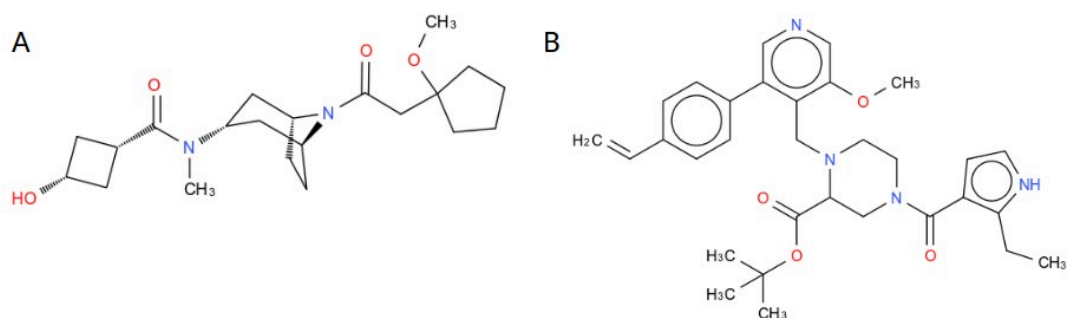


Figure 32: Molecules corresponding to the data points highlighted by the letters in the PMI plot in Figure 31. (A) Natural product-like compounds at the edge between the disc- and rod-like shape. (B) Library member tending to sphere-like shape.

6.5 Conclusions

This chapter introduces a novel and easily applicable method for assessing building blocks for DEL design, using the KNIME platform. Since the libraries did not target any specific protein, diversity was the predominant criteria for the selection to increase the probabilities of *hit* identification. It is worth mentioning that diversity was intended in terms of chemical properties as well as geometrical shape. The building blocks were selected according to the chemistries that could potentially be utilized on DNA tagged substrates. For the first time, a library based on the Povarov, Biginelli and *aza*-Diels-Alder reactions was designed providing an access to the tetrahydroquinoline, dihydro-pyrimidinone and dihydro-pyridinone scaffolds, respectively. In the second part, the Suzuki coupling was the elected reactions coupled with a preliminary reductive amination. Based on the selected building blocks, the virtual libraries were enumerated and appropriately labelled. The molecules resulting from the enumeration displayed high diversity, which confirmed the assumption that diverse building block selection ensure diverse final products. The libraries members were compared with available databases of drug-like compounds and, especially the subsets based on the *aza*-Diels-Alder and the Suzuki scaffolds, showed high similarity with drug-like compounds both in terms of chemical properties and shape. In summary, these chemistries constitute valid options for practical library synthesis.

7 Hits validation by molecular docking

7.1 Introduction

7.1.1 Molecular docking for *hit* identification

In drug discovery, a large part of new compounds derive from screening campaigns such as high throughput screening (HTS), which supported hit identification for years.[120] However, nowadays larger library sizes are required and those methods revealed to be inadequate for the task. [2] Larger libraries allow for wider investigation of the chemical space and by consequence may increase the possibilities to uncover novel binders. DNA-encoded libraries (DELs), due to their combinatorial nature, allow for exponential library growth and render the process of library synthesis more suitable for ultra-large screening campaigns. Therefore they constitute a valid alternative to screening of discrete compounds libraries.

A second option for *hits* identification are computational methods such as *docking*. A *hit* molecule is defined as the chemical entity that shows high activity or affinity in screening campaigns, but it needs to be optimized via the *hit-to-lead* procedure. The docking protocol, which could be considered as a *virtual screening*, consists of two phases, the *posing* and the *scoring*. The posing is the prediction of the conformations assumed by the ligand in the binding pocket of the target proteins. [121] The posing performs a conformational search on the ligand within the binding site of the target protein by varying parameters such as rotatable bonds coordinates or bonds length, until the energy converges to a minimum (Figure 33). In the convergence condition, a conformational change does not affect the calculated energy.

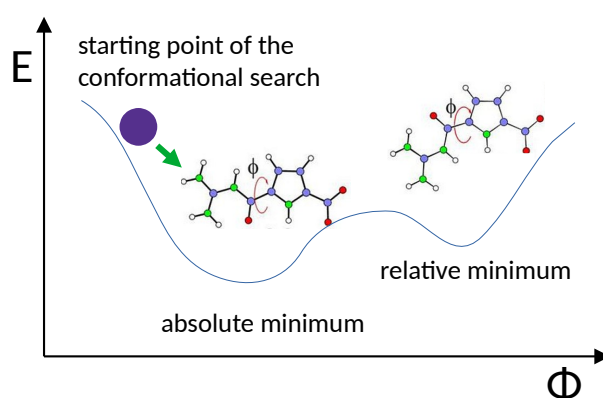


Figure 33: Example of conformational search by varying torsion angles Φ . The energy E changes in function of the torsion angle Φ and, according to the starting point, an absolute or relative minimum can be reached.

The *scoring* part consists of ranking the conformations according to the calculated energy which gives information about the affinity between the protein and the ligand. The scoring is calculated via *scoring functions* (SFs) which can be divided in different categories: physics-based, empirical, knowledge-based and machine learning-based. [122] The difference between them lies in the type of regression method utilized for predicting the energy of the conformation: the first three rely on linear regression, whereas the fourth on nonlinear regression, based on machine learning algorithms. For the scope of this thesis, only the first class of scoring functions is considered. The force fields-based SFs calculate the binding energy as enthalpy contribution, including the van der Waals and electrostatic potentials. [123] Force fields (FFs) are approximations used to calculate the enthalpic contribution to the *Gibbs free energy* of the protein-ligand complex. An exemplary function for a force field is illustrated in Equation 3. [124]

$$E = \sum E_{bonds} + \sum E_{angles} + \sum E_{dihedral} + \sum E_{non-bonded} \quad (\text{Eq. 3})$$

E_{bonds} , E_{angles} and $E_{dihedral}$ represent the energies of bonds, angles and dihedral (three-dimensional angles formed at the intersection of planes), respectively. The $E_{non-bonded}$ parameter is related to Coulomb and van der Waals interactions. This calculation is repeated for all atoms of the molecules every time a conformational change occurs. The conformation that gives the overall minimal energy is considered to be the most realistic.

Since the FFs-based scoring functions omit the entropy and the solvent effect, they were soon improved by including solvation models in the calculation, therefore increasing the reliability of the results. [125] The previous equation could be approximated to the Equation 4, where an additional parameter accounts for the hydration contribution.

$$E = \sum E_{bonds} + \sum E_{angles} + \sum E_{dihedral} + \sum E_{nonbonded} + \sum E\sigma\Delta SA \quad (\text{Eq. 4})$$

Let σ be the solvation parameters and ΔSA the difference in SASA (solvent accessible surface area) between the unbound and the bound states. [126]

In the pose calculations, two levels of accuracy have been achieved over time. The earliest approach is the rigid docking and it is mostly based on the key-lock assumption. The ligand is searched within six degrees of freedom and translational changes. [127.] In this approach, the changes in energy stemming from the binding event are ignored, yet later, the flexible docking was developed to increase the accuracy. This protocol also considers the conformational changes of the protein backbone which

also influence the position of many residues. [128] Although most of the available docking software nowadays can validate experimental data and therefore are considerably reliable, the flexibility of the target proteins remains challenging to predict.

Docking algorithms are based on combinations of the aforementioned features, so they are characterized by distinctive assumptions and approximations. In this work, two algorithms in particular are introduced: *Hyde*, as part of the SeeSAR package developed by BiosolveIT [129] in collaboration with the group of Prof. Rarey at the university of Hamburg, and *Glide*, belonging to the Schrödinger suite. [130]

Hyde calculates the affinity between the ligand and the protein as a function of the difference in energy due to the displacement of water molecules surrounding the ligand and filling the binding pocket of the protein during the binding event. Moreover, it considers the increase in energy due to the favourable hydrophobic interactions between the ligand and hydrophobic aminoacid residues of the protein that would otherwise be solvent-exposed. Docking calculations performed by *Hyde* result in the pose of the ligand into the binding pocket, in which atoms are surrounded by coloured spheres according to their effect on the interactions. If the specific atom's interactions are considered favourable, the sphere is green, otherwise red. Additionally, the affinity is expressed in terms of concentration to assist those who are not familiar with the concept, to interpret the docking results. [131]

Glide is based on force fields, therefore, the scoring function calculates the energy of each atom of the ligand, including its torsion energies, the electrostatic and van der Waals potentials, and additionally, the solvation model. Although this may appear highly demanding in terms of computational resources and time, *Glide* performs a funnel-like search which become more stringent only when the possible conformations are reduced by preliminary selection phases. In fact, *Glide* presents two levels of accuracy: the “standard precision”, a softer sampling of conformational space, and the “extra precision”, more complex calculations considering all the aforementioned factors. The result of this docking protocol are poses with the respective scores representing the difference in binding energy (ΔG_{bind}). Therefore, the lower the score, the lower the energy of the complex is, compared to the unbound ligand. In other words, poses with lower docking scores are more reliable as *hits*. [132]

7.1.2 Critical steps in a docking procedure

7.1.2.1 Ligands preparation

In this step, all possible ligand conformations are generated by varying all torsion angles sequentially. Additionally, the hydrogenation patterns and charges are applied according to the pH.[3]

7.1.2.2 Protein preparation

Prior to any calculation the protein structure must be chosen and prepared. Generally, structures with high resolution are to be preferred and the *holo* (complexed) state performs better than the *apo* (free) state in docking experiments. The preparation is necessary to overcome artifacts due to the crystallization process such as mutations, buffer components such as PEG, cofactors and missing loops or side chains. Moreover, the protonation pattern of the protein has to be defined newly since hydrogen atoms are usually not well resolved in crystal structures. This aspect is particularly important for residues that might be charged such as Glutamate and Aspartate, as their protonation state impacts coulomb interactions and hydrogen bonds formation. Finally, the conformation of the protein has to be optimized by energy minimization since during the crystallization process, the protein solution is highly concentrated to permit the crystallization. However, this situation does not reflect physiological conditions and due to the concentration, the protein may assume unnatural conformations. [3]

All this applies when an experimental structure of the protein is available, otherwise *Homology modeling* can be employed to generate plausible structures according to preexisting three-dimensional structures with high similarity. [133] The most noteworthy software for *Homology modeling* is *AlphaFold*, which outperformed other options as demonstrated by J. P. Roney and S. Ovchinnikov [134]. Nevertheless, some limitations still remain, especially related to disordered domains, [135] so scientists prefer to rely on experimental structures.

7.1.2.3 Definition of the binding site

In the case of unknown binding site and interaction mode for a resolved structure, the computational software can help with exploring the protein surface for druggable pockets. Those areas are generally unoccupied and slightly buried surfaces that could potentially host a ligand. Moreover, they should contain specific pharmacophoric features that allow the interaction with an external molecule. An example of such functions is the protocol *SiteMap* implemented in Maestro as part of the Schrödinger

suite. This methodology evaluates the size and the openness of the pocket as well as the hydrophobic/hydrophilic character. The openness of the pocket here is related to the exposition of the aminoacid residues to the solvent. If multiple pockets are detected, a score is assigned according to the above mentioned criteria. After separating the ligand from the protein, the *SiteMap* protocol was applied and was able to identify the same binding site as the initial ligand in most of the experiments.

[136]

7.1.3 DEL and docking

To increase the confidence in *hit* identification, DNA-encoded libraries and docking might be used in series. The docking in this case would function as validation methodology to counterbalance some experimental artefacts due to the presence of the DNA tag and the set up of the DEL screening. In fact, the calculation of the EFs after the selection assay can be influenced by unspecific interactions with beads or by the protein denaturation due to the buffer or other conditions. In this way, the docking would aid in the identification of eventual false positives, i.e. compounds that, despite their high EF (enrichment factor) result to be inactive in further biological assays. Additionally, the visual inspection of docking poses and the analysis of the scores might hint to possible modification for the *hit-to-lead* optimization process. Thus, not only the identified hit is to be submitted to organic synthesis but also congeneric series. This strategy would increase the chance of success.[137]

Alternatively, the docking could be utilized in parallel or even prior to the library synthesis and screening. If used in parallel, the docking could detect false negatives, i.e. molecules that resulted good binders in the docking procedure but not in the selection assay. By consequence, the investigation of those molecules might serve the improvement of the DEL technology. In fact, after the synthesis or purchase of those molecules from big databases, [108],[109] biological tests might follow. If the tests confirm the activity, the reasons behind the lack of affinity in the selection assay could be inspected. Similar to the previous case, also in this case the docking might serve the *hit* optimization process. [138] In summary, the combining the DEL screening and the docking can be more effective in identifying active compounds than using them separately.

7.1.4 Protein selection in DEL

In a DEL screening campaign, as well as in docking experiments, the choice of the proteins to assess is a crucial criterion to ensure success and novelty. In general, for untargeted libraries proteins belonging to distinct families are selected according to medical need or to assess *ligandability*. Soluble proteins

without known ligands are preferred, yet unexplored proteins represent more attractive targets to find original ligands. However, the challenging membrane proteins have become a viable options as well. [139] Additionally, for the selection assay in which the protein is bound to a solid support and the library is in solution, the proteins must tolerate to be modified with Histidine or Glutathione tags to bind the beads. Even before the protein meet all those requirements, its biological significance is to be estimated. Cancer-related proteins acquired high interest in the latest years so they represent valid target choice.[140] Special attention is directed towards proteins controlling the apoptotic cycle, such as the Bcl 2 (B cell lymphoma 2) family members and the MDM2 (murine double minute 2) protein, or the cellular growth such as MKK7 (mitogen-activated protein kinase kinase 7).

The Bcl 2 family is characterized by two classes of proteins: pro-survival, such as BAK and BAX, and pro-apoptotic, such as BCL-2 and BCL-X_L. In case of cellular stress, the pro-apoptotic proteins oligomerise and provoke mitochondrial outer membrane permeabilization (MOMP), whereas the pro-survival proteins explicate their function by sequestering the pro-apoptotic proteins (Figure 34). The delicate regulation of this phenomenon plays a role in the balance between cell death and cancer. [141]

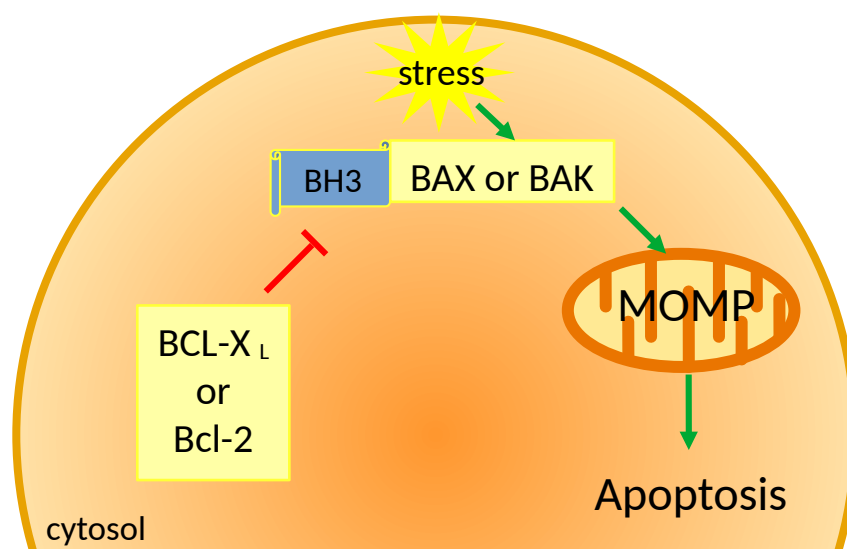


Figure 34: BCL-X_L and BCL-2 mechanism in sequestering the pro-apoptotic proteins BAX and BAK, through the BH3 domain. In case of stress, the BAX and BAK would oligomerize and provoke the mitochondrial outer membrane permeabilization (MOMP), which initialize the apoptotic process.

The interaction between pro-survival and pro-apoptotic proteins takes place in correspondence of the BH3 domain (Figure 35 A), which has been therefore used as model for potential inhibitors of this interaction (Figure 35 B).

The resulting compound *ABT-737* proved to be effective against solid tumors [144] This discovery paved the way for other compounds targeting the BCL-X_L protein which revealed to be active (Figure 36). [145], [146].

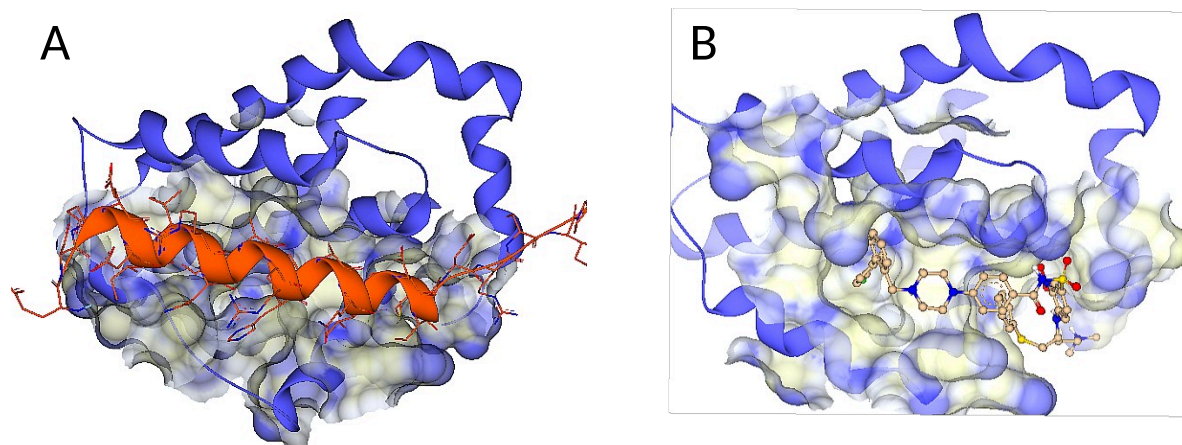


Figure 35: Three-dimensional structure of BCL-XL in complex with (A) the BH3 domain of a proapoptotic protein (PDB: 4QVE [142]) and (B) the ABT-737 ligand (PDB: 2YXJ [143]). ABT-737 occupies the same pocket as the BH3 domain.

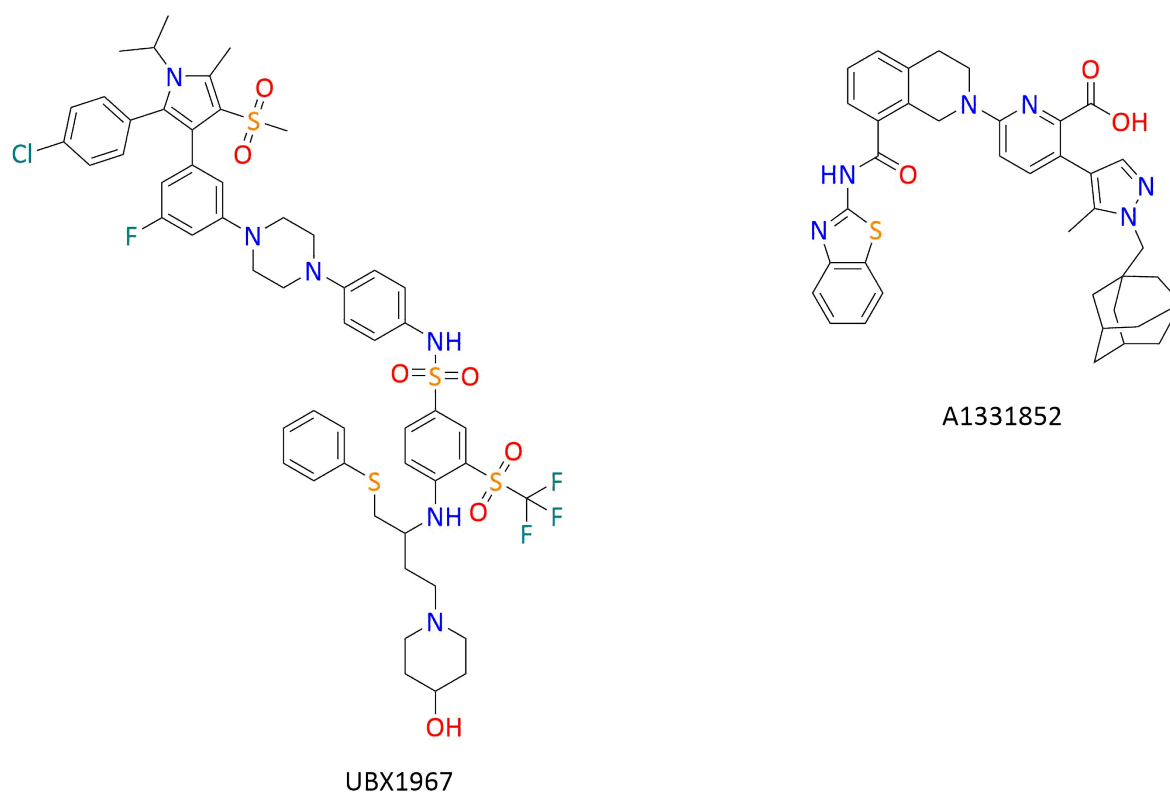


Figure 36: Drug candidates for targeting the BCL-X_L/BH3 domain interaction. [145], [146]

MDM2 is the ubiquitin ligase E3 which promotes the degradation of p53 in normal cells by ubiquitination. The transcription factor p53, called “the guardian of the genome”, is a potent tumor suppressor as it induces cell cycle arrest in conditions of stress or genomic instability (Figure 37). [147] The inhibition of this PPI (protein-protein interaction) (Figure 38 A) leads to higher production of the p53 protein, which has a beneficial effect against tumour cells.

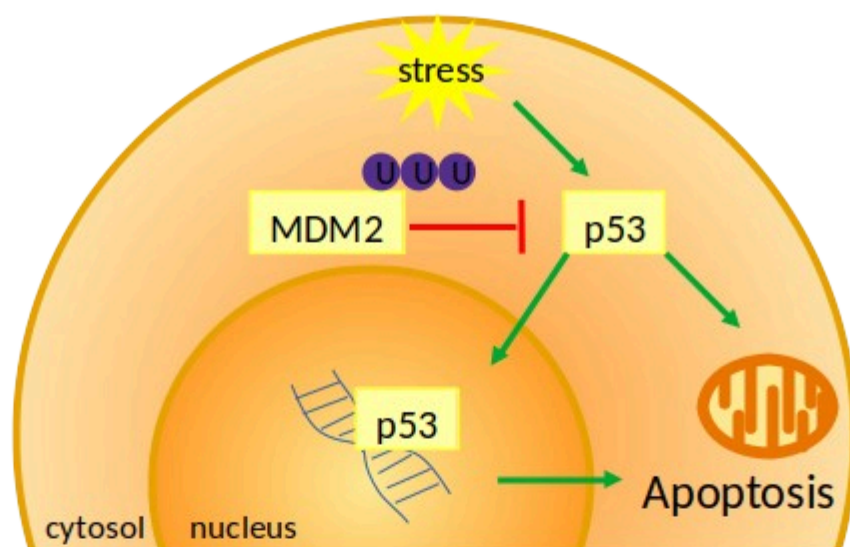


Figure 37: MDM2-p53 regulation mechanism. MDM2 ubiquitinates the p53 protein, that in case of stress initiates the apoptotic process. The ubiquitination of p53 lead to its degradation, cell survival and proliferation.

The discovery of the activity of Nutlins (Figure 38 B) against this target [148] enabled the exploration of multiple inhibitors which are currently under clinical trials (Figure 39). [149]

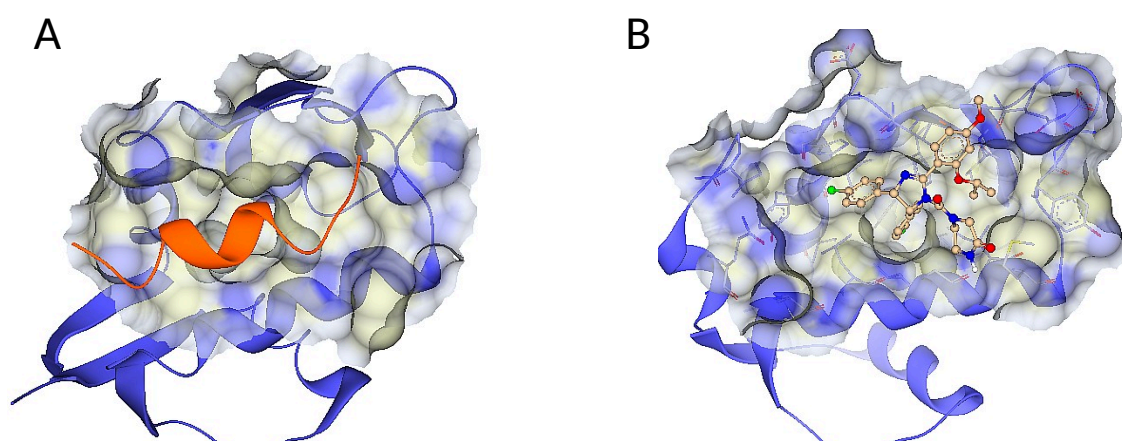


Figure 38: Three-dimensional visualization of the MDM2 protein in complex with (A) the p53 domain (PDB: 3V3B) [150] and (B) a member of the Nutlin family (PDB:4HG7). The Nutlin occupies the same pocket as the p53 domain. [148]

MKK7, also known as MEK7 or MAP2K7, is part of the kinase signaling pathway which control cell growth, proliferation and apoptosis. In particular, it activates the c-Jun N-terminal kinase (JNK) by threonine-phosphorilation (Figure 40). The functional contrariety of JNK is still under research and it involves regulation of responses against stress and inflammation towards either apoptosis or cell proliferation according to the distinct downstream substrates. [151]

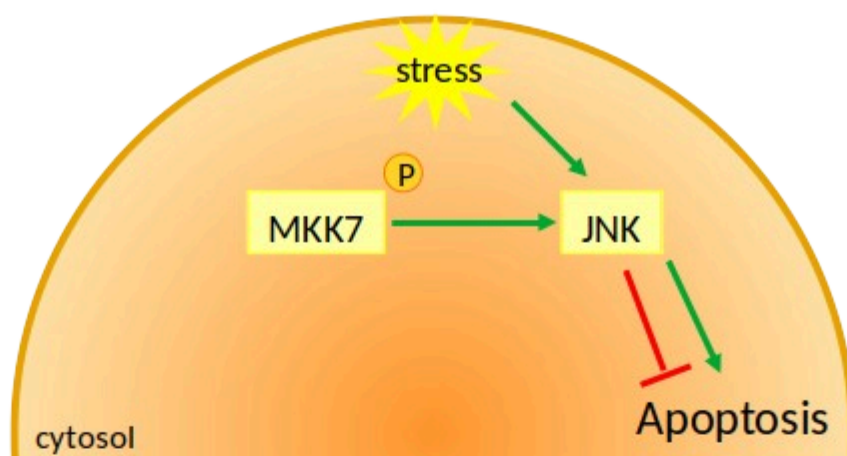


Figure 40: MKK7 mechanism. The kinase activates by Threonine phosphorilation the protein JNK which according to the downstream substrates can induce or repress apoptosis.

The MKK7 protein seems to activate the JNK towards the latter direction, so inhibitors against this target have been researched. The challenge concerning kinase inhibitors is that they target the ATP-site, called *hinge* domain, which is highly conserved among all kinase families, making the discovery of selective kinase inhibitors challenging. This challenge may be overcome by designing covalent inhibitors to non-conserved Cysteine residues in the active site. In particular, a successful strategy sees the covalent bond between the acrylamide moiety of such inhibitors and the sulphur of Cysteine 218 (Figure 41 and 42). [152], [153]

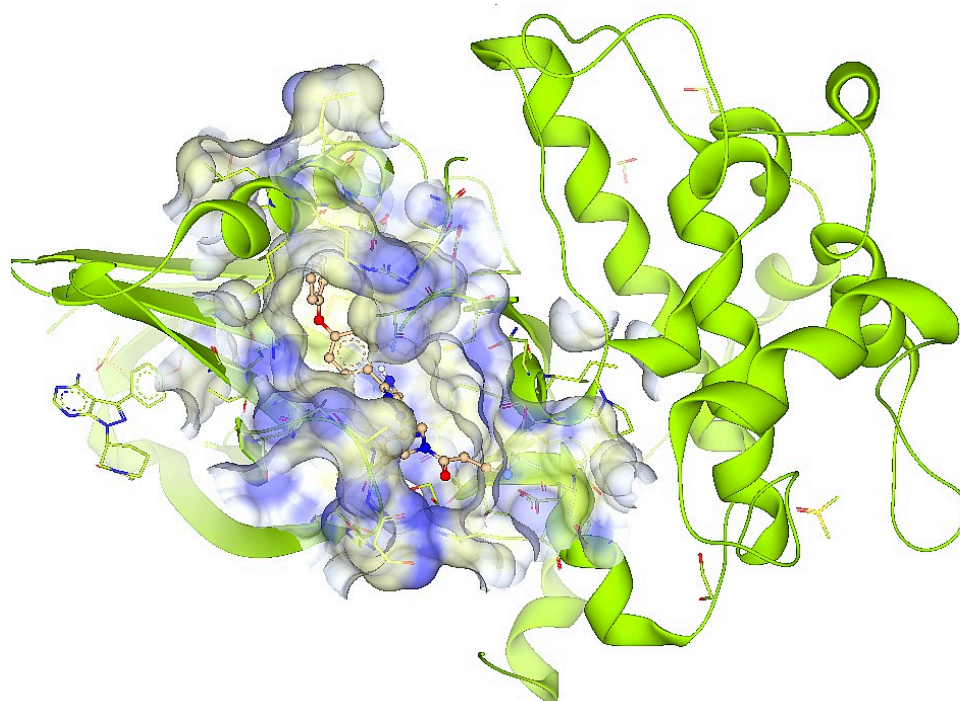


Figure 41: Three-dimensional structure of MKK7 in complex with the covalent inhibitor Ibrutinib. Interestingly, the same molecules reversibly interacts with an allosteric site, beside binding to the ATP site of the protein. PDB: 6YG2. [154]

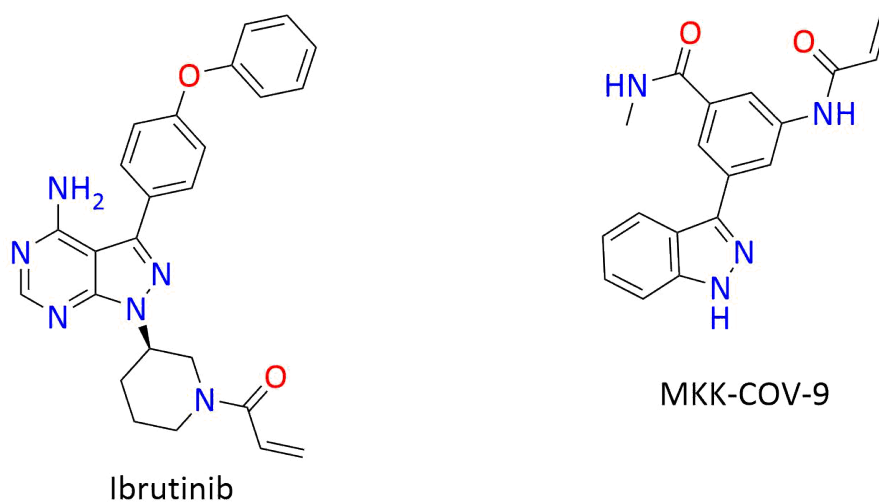


Figure 42: Drug candidates covalently targeting the ATP binding site of MKK7.[152], [153]

For their role in cancer research and their affiliation to very different families, those three proteins constitute a useful set for validating the DEL technology with the aid of molecular docking.

7.2 Aim

DNA-encoded libraries selection assays allow for panning huge numbers of molecules against multiple proteins at the same time and this is made possible by the unique DNA tag coupled to each library member. The data load generated by the selection assay and the following sequencing protocols could result overwhelming if tackled without the proper tools. Moreover, due to different factors that may affect the reliability of the selections assay, a proof of validation is needed, to reduce noise and to increment the confidence in *hit* identification. An established procedure for DEL validation and *hit* selection is currently missing and this is the purpose of this part of the thesis. In particular, the focus is put on chemoinformatics and computational methods that aid in selecting the more significant molecules that are worth synthesizing for biological tests, discarding compounds that could have unspecific interactions and may reveal inactive.

7.3 Methods

In our laboratories, a 100,000-membered library was synthesized and screened against 16 targets. The library was based on three multi-component reactions (MCRs), namely Povarov, SnAP and Ugi-aza-Wittig, forming tetrahydroquinoline (THQ), piperidine and oxadiazole scaffolds respectively. MCRs are the most suitable tool for DEL synthesis as they provide high structural diversity in only one step. The 16 screened proteins range over many different target classes, from kinases to protein-protein interaction (PPI) targets. After the selection assay and following sequencing, in order to prioritize the molecules for resynthesis off-DNA, the enrichment factors (EFs) for each molecule were calculated. The EF were intended as the frequency of each sequence per protein (rank abundance, RA) after the selection assay normalized over the same count in the native DEL before the selection experiment (Eq. 5). [11]

$$EF = \frac{RA_{selection}}{RA_{native\ DEL}} \quad (\text{Eq. 5})$$

The KNIME workflow for prioritizing *hits* was applied on the 16 proteins screened during the selection assay but only MDM2, BCL-X_L and MKK7 were reported within the scope of this thesis. Those proteins were considered a good validation as they belong to very different families.

Initially a so-called *Structure sheet* file, containing all the building blocks and the respective sequences, was employed to track back the substructures and assemble them into the final molecules. The control experiments, such as the selection against empty beads, were excluded from the analysis as not interesting for prioritizing hits and only molecules with EF higher than the median for each protein were considered for further processing. As a first step, the frequent hitters were excluded, defined as molecules which interacted with more than four proteins. This number depended on the composition of the set of screened proteins. In fact, it was plausible that molecules such as kinase binders would show a high EF for all the four members of this target class screened in our assay. In this way, we reduced to some extent the chance of privileging false positives, which is in general quite challenging in combinatorial libraries screening assays. Hence, the remaining combinations of cycle 1, 2 and 3 molecules were labeled in order to identify them after the docking procedure and the fragments were assembled within the KNIME Analytics

Platform. Although the library was based on MCRs, the virtual synthesis was split into sequential steps. Firstly, the connection to the DNA was substituted by a short linker via an amide bond formation, because a longer linker would greatly affect the binding affinity of the molecules for the protein. The interference of the DNA tag on the affinity between the molecule and the target could be objected for the library screening as well. Nevertheless, the absence of a long linker is compatible with the logic of *hit* re-synthesis, as molecules with a long linker would hardly show drug-like properties. The final molecules were then virtually synthesized for each proteins in an automatic fashion. With the virtual molecules in hand, the docking simulations were performed using two different algorithms in order to confirm or contradict the EFs calculation. The additional function of the docking in our hands was to suggest modifications that could potentially increase the affinity between the compounds and the target proteins. This would improve the process of *hit-to-lead* optimization in terms of time and effectiveness, as instead of the single *hit*, series of derivatives could be synthesized. The first algorithm was *SeeSAR*, trademark of the *BiosolveIT*. After preparing the protein and ligands with the respective functions implemented in *SeeSAR*, the docking protocol was carried out. In the preparation of the protein the co-crystallized ligand as well as other species were eliminated in order to not bias the docking towards specific pockets. This strategy was named *blind docking*, as the preparation of the proteins was not affected by any additional ligand and the docking poses were calculated without any constraints. The *blind docking* was applied because the molecules resulting from the DEL are usually different from co-crystallized ligands, so constraints based on known binding mode would not help in the *hits* identification. Moreover, we did not want to force the compounds to follow a specific binding pattern as this library was not conceived as target-based and the binding mode was unknown. In fact, we employed the docking for this reason as well: if after resynthesis, one hit confirmed to be active, the docking algorithm could suggest the most probable pose. Therefore, the druggable pockets were identified by the software and the docking was performed over all of them. A similar procedure was employed with the second docking program *Maestro* from the *Schrödinger* suite, whose docking protocol is called *Glide*. Similar to the docking with *SeeSAR*, the proteins were prepared after removing all additive species, the binding pockets were identified via the *SiteMap* protocol and the *blind docking* was performed.

7.4 Results and discussion

The sequencing data were firstly submitted to EFs calculation which was performed by Nils Jannick Schüssler, master student in Prof. Fried group at the department of statistics at the TU Dortmund. The final output table contained information about each combination of building blocks per each protein, as exemplified by the extract in Table 16. As the calculation was performed for every protein, the first column referred to the protein number. Each row represent a combination of building blocks which were characterized by the IDs in the second, third and fourth columns. The EF column included the EF values, normalized over the overall abundance (*rank abundance*) as exemplified in the *Methods* section, and the last column contained information about the control experiments utilized as validation of the sequencing protocol. Since this table did not provide any information neither about the molecular structure of the final molecules nor about the *hit* identification, it was processed with the software KNIME.

Table 16: Extract of the initial table with the EF calculation.

























Protein entry	cycle 1 ID	cycle 2 ID	cycle 3 ID	enrichment factor	rank abundance	control experiment
1	Aldehyde15	Scaffold3	Acid138	1	74744.5	Streptavidin beads
1	Aldehyde15	Scaffold32	Acid100	0	75913.0	Streptavidin beads
1	Aldehyde15	Scaffold4	Acid15	3	69233.5	Streptavidin beads
1	Aldehyde15	Scaffold45	Acid148	1	73248.0	Streptavidin beads
1	Aldehyde2	Scaffold1	Acid32	1	34481.0	Streptavidin beads
1	Aldehyde2	Scaffold1	Acid37	27	67680.5	Streptavidin beads
1	Aldehyde2	Scaffold11	Acid37	9	77468.0	Streptavidin beads
1	Aldehyde2	Scaffold2	Acid37	1276	79013.0	Streptavidin beads
1	Aldehyde2	Scaffold24	Acid37	43244	79023.0	Streptavidin beads
1	Aldehyde2	Scaffold3	Acid137	6	51079.5	Streptavidin beads
1	Aldehyde3	Scaffold1	Acid127	0	78884.0	Streptavidin beads
1	Aldehyde3	Scaffold1	Acid132	0	78966.0	Streptavidin beads
1	Aldehyde3	Scaffold1	Acid37	0	72612.5	Streptavidin beads
1	Aldehyde3	Scaffold2	Acid127	2	70881.0	Streptavidin beads
1	Aldehyde3	Scaffold2	Acid37	218	78879.0	Streptavidin beads
1	Aldehyde3	Scaffold2	Acid43	2	76655.0	Streptavidin beads
1	Aldehyde3	Scaffold24	Acid37	11	76304.5	Streptavidin beads

7.4.1 KNIME workflow

In order to visualize the structures and to generate the corresponding molecules, this resulting table was combined with the initial file that was used in designing the library, in which each BB ID was related to the corresponding SMILES. Moreover, the protein number was joined to the actual name of the targets present in the same file. After this essential step, the resulting table was more

interpretable and an extract of randomly picked entries is depicted in Table 17. Here, the BB IDs are coupled with the respective chemical structure in the sixth, seventh and eighth columns and the protein name is appended as well. It is worth mentioning that the "cycle 2 SMILES" column represent the whole scaffolds of the molecules that could be based on the Povarov, Ugi-*aza*-Wittig or SnAP reactions. However, the IDs depended on the *synthons* utilized in the reactions so they are unique like the others.

Table 17: Resulting table after combining the BBs IDs with the respective molecular structure and the protein ID number with the respective name.

Protein entry	cycle 1 ID	cycle 2 ID	cycle 3 ID	EF	cycle 1 SMILES	cycle 2 SMILES	cycle 3 SMILES	Protein name
8	Aldehyde2	Scaffold22	Acid38	358				MKK7
12	Aldehyde11	Scaffold24	Acid37	155				MDM2
12	Aldehyde4	Scaffold1	Acid9	18				MDM2
13	Aldehyde4	Scaffold1	Acid107	672				MDM2
13	Aldehyde11	Scaffold24	Acid37	88				MDM2
13	Aldehyde6	Scaffold24	Acid37	61				MDM2
14	Aldehyde10	Scaffold3	Acid19	312				BCL-XL
14	Aldehyde3	Scaffold24	Acid19	239				BCL-XL

7.4.1.1 Frequent hitters identification

Similar to other screening technology, DEL suffers the issue of false positives, defined as highly enriched molecules that reveal to not bind the target protein in biological assays. Our preliminary solution to face this problem was to exclude molecules binding more than four target proteins. The first reason for this was that even if they were actual binders, they would have been unspecific and therefore not attractive for optimization. The second reason is related to the

technology itself, as highly enriched molecules in most screening experiments might bind not to the proteins but to the beads or other materials used in the assay. In our evaluation, nine combinations were therefore excluded from further analysis. In Table 18, these nine combinations are listed as rows, displaying the BB IDs and the respective chemical structures. The last column refers to the sum of proteins that interacted with each of those building blocks combinations. It is noticeable that beside combinations binding five or six proteins, the combination between aldehyde 15, scaffold 24 and acid 37 formed interactions with 13 proteins, which accounted for almost all screened proteins.

Table 18: Frequent hitters.

cycle 1 ID	cycle 2 ID	cycle 3 ID	cycle 1 SMILES	cycle 2 SMILES	cycle 3 SMILES	protein count
Aldehyde10	Scaffold24	Acid37				6
Aldehyde11	Scaffold24	Acid37				5
Aldehyde15	Scaffold2	Acid37				9
Aldehyde15	Scaffold24	Acid37				13
Aldehyde2	Scaffold24	Acid37				7
Aldehyde3	Scaffold2	Acid37				5
Aldehyde4	Scaffold24	Acid37				8
Aldehyde9	Scaffold2	Acid37				7
Aldehyde9	Scaffold24	Acid37				5

Notably, the combinations, and not the single synthons, were considered as frequent hitters not to narrow down too much the molecules scope and risk to exclude actual binders. Moreover, the final library members should be composed by all three cycles building blocks. Although this assumption cannot be proofed during the library synthesis due the *split&pool* technology, it is expected that all parts of the molecule account for the interactions. For this reason, the combinations of the three cycles BBs were considered and counted.

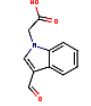
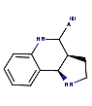
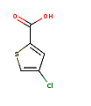
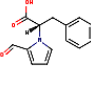
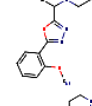
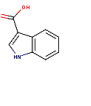
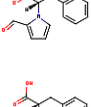
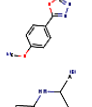
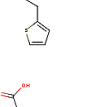
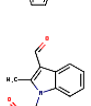
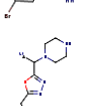
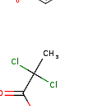
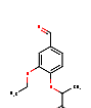
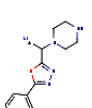
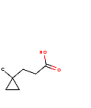
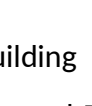
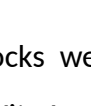
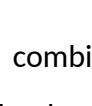
Overall, specific structures could be recognized in the case of cycle 2 and cycle 3 building blocks, while the cycle 1 aldehyde showed more variety. In particular for cycle 2, scaffold2 and scaffold24 appeared in all frequent hitters. Scaffold 2 was based on the Povarov reaction and it was characterized by the 2-ethyl aniline, for which the lipophilic character most likely played a role in binding the polypropylene of the tubes. Scaffold 24 was based on the Ugi-*aza*-Wittig reaction and it was constituted by the 4-bromo-phenyl carboxylic acid whose unpolar character might have affected the interaction with many target pockets. For cycle 3, the BB acid37 prevailed overall, most likely due to unspecific binding with the magnetic beads. In general, such structures are to be avoided in library design and can be filtered *a priori* by chemoinformatics tools. Finally, the most frequent cycle 1 fragment was aldehyde 15, followed by other six aldehydes: 4, 2, 9, 10, 11 and 3. A structural similarity can be identified between aldehyde 15, 10 and 11, sharing the 2-phenoxypropionic acid moiety, and between aldehyde 9 and aldehyde 4 characterized by the indole substructure binding up to eight targets. The combination of aldehyde 15, scaffold 24 and acid 37 was above all the most frequent, binding almost all the considered proteins (13 out of 16). The exchange of the cycle 2 structure with scaffold24 produced the second most common combination binding 9 out of 16 proteins. It is noteworthy that the single synthons were still present in the dataset but within different combinations.

7.4.1.2 Molecule generation and labelling

For the purpose of docking we needed to enumerate the selected *hits* based on the combination of building blocks and, in order to track the docked poses to the specific combinations, one essential step in this directions was to label them beforehand. An extract of randomly picked entries in the table after the labelling step is illustrated in Table 19. To each building block

combination an ID was assigned, composed by the prefix "comb" and a numeric sequence generated automatically (last column in the table).

Table 19: Examples of labelled combinations.

Protein entry	cycle 1 ID	cycle 2 ID	cycle 3 ID	EF	RA	cycle 1 SMILES	cycle 2 SMILES	cycle 3 SMILES	protein name	combination ID
12	Aldehyde 14	Scaffold 1	Acid 9	958	75733				MDM2	comb_848
12	Aldehyde 1	Scaffold 29	Acid 106	657	76711.5				MDM2	comb_854
14	Aldehyde 1	Scaffold 17	Acid 19	1172	78276.5				BCL-XL	comb_1004
14	Aldehyde 1	Scaffold 9	Acid 38	816	78684.5				BCL-XL	comb_1009
8	Aldehyde 12	Scaffold 28	Acid 75	2306	78713.5				MKK7	comb_535
8	Aldehyde 10	Scaffold 21	Acid 49	1056	78684.5				MKK7	comb_554

After the labelling step, the cycle 1, 2 and 3 building blocks were combined employing the chemistry-oriented extensions in KNIME, ChemAxon and RDKit. A randomly picked extract of the final outcome is depicted in Table20.

Table 20: Final products of the library after the synthetic steps.

EF	cycle 1 SMILES	cycle 2 SMILES	cycle 3 SMILES	protein name	combination ID	product
1701				MDM2	comb_841	
852				MDM2	comb_850	
1908				BCL-XL	comb_999	
6443				MKK7	comb_527	
4367				MKK7	comb_530	
1984				MKK7	comb_537	
1056				MKK7	comb_554	

In the final structure of the products (seventh column in the table), the carboxylic acid moiety was transformed into an amide group. This approximation needs to be considered in comparing the selection assay and the docking results, as the linker and the DNA tag might play a role in the affinity during the selection assay, while this effect is ignored during the docking procedure. *In silico* experiments involving a linker containing a long unsaturated hydrocarbon chain (where the

DNA tag in attached in the DNA-encoded molecule) demonstrated that there were interactions between the linker's amide bond and aminoacid residues in the binding site, while the long unsaturated chain tended to occupy hydrophobic pockets on the protein surface (Figure43). For example, in the case of the protein BCL-X_L, the binding pocket is formed by two sub-pockets and one of them is relatively hydrophobic,[143] so in this case the presence of the linker occupying the second sub-pocket greatly increased the calculated affinity between the molecule and the protein. Although such a linker mimics the steric hindrance of the DNA and therefore better represents the experimental conditions in a DEL selection assay, such molecules would be irrelevant from the point of view of *hit*-resynthesis and *hit-to-lead* optimization.

Therefore, we decided to disregard the DNA/linker effect on the affinity calculations and the following experiments were performed with a short methyl amido linker which influenced the ligand-protein interaction to a lesser extent. The carbonyl moiety of the amide group, acting as hydrogen bond acceptor, affected to some extent the affinity calculation anyway but this was considered irrelevant.

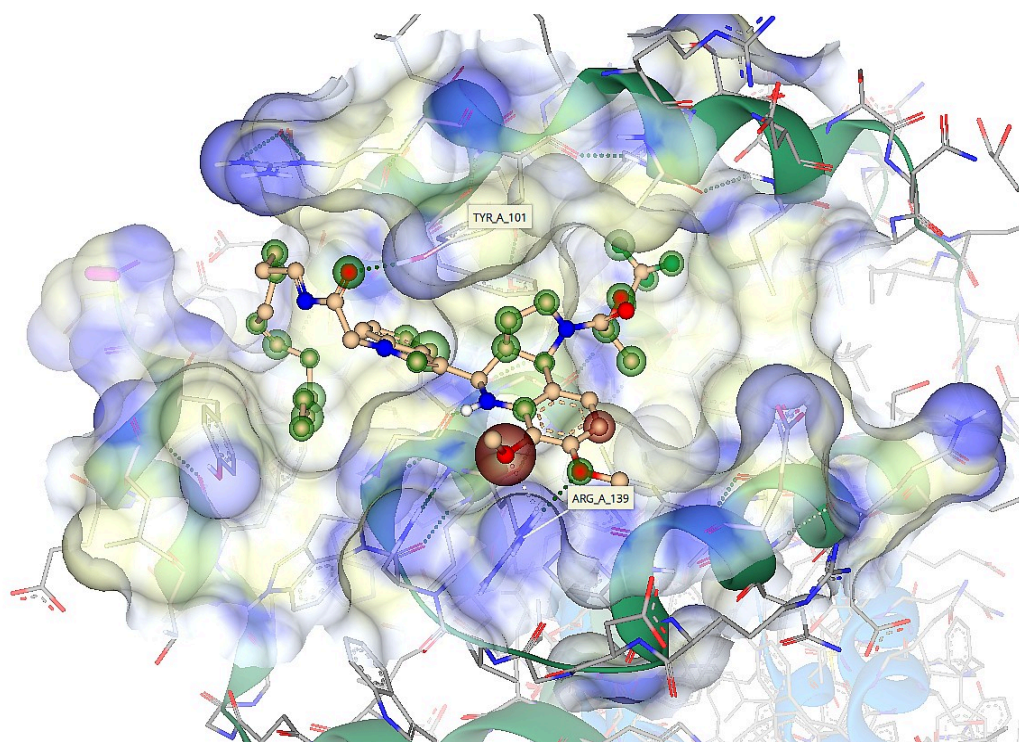


Figure 43: Position of linker with a long and unsaturated chain inside the pocket of the BCL-X_L protein. Figure generate by SeeSAR.

7.4.1.3 Correlations between building blocks and proteins

After generating all ligands per each protein, some trends were observed regarding the preference of some proteins towards specific building blocks. The cycle 1 or cycle 3 fragments were plotted in the *heat maps* depicted in Figure 44 and 45.

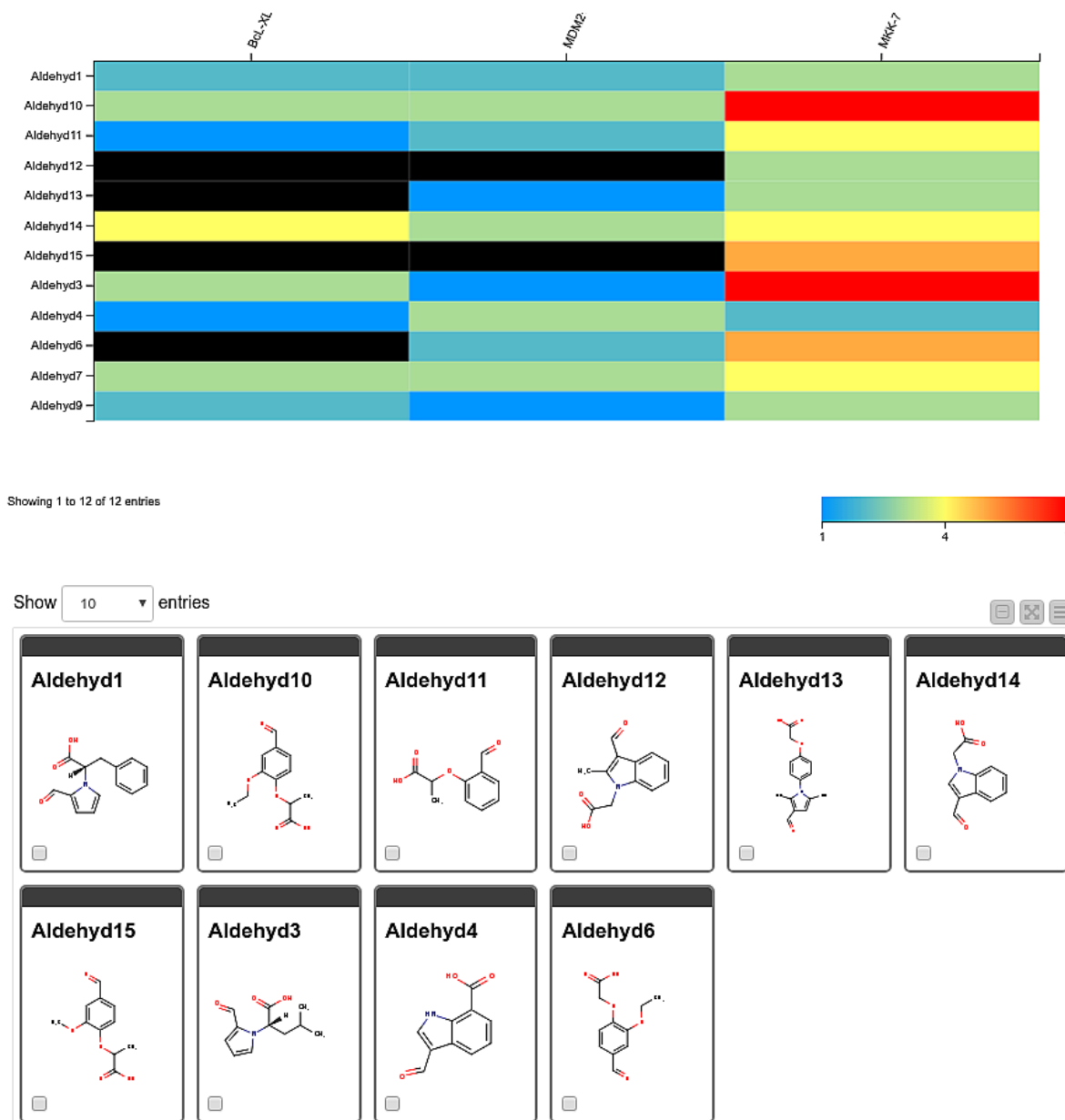


Figure 44: Correlations between the cycle 1 building blocks and proteins. The cells of the heat map are color-coded according to the count of molecules presenting the specific substructure for each of the three considered proteins.

Among the cycle 1 aldehyde building blocks, a clear pattern was in general not detectable, except for target MKK7 towards Aldehyde3 and 10. In Figure 44, the red cell at the intersection between

the column "MKK7" and the two rows corresponding to Aldehyde 10 and 3 highlights this correlation. In contrast, MDM2 bound predominantly the indole-based acid Acid 106 and BCL-X_L tended towards the 2-thien-2-ylpropionic acid Acid 19 (Figure 45).

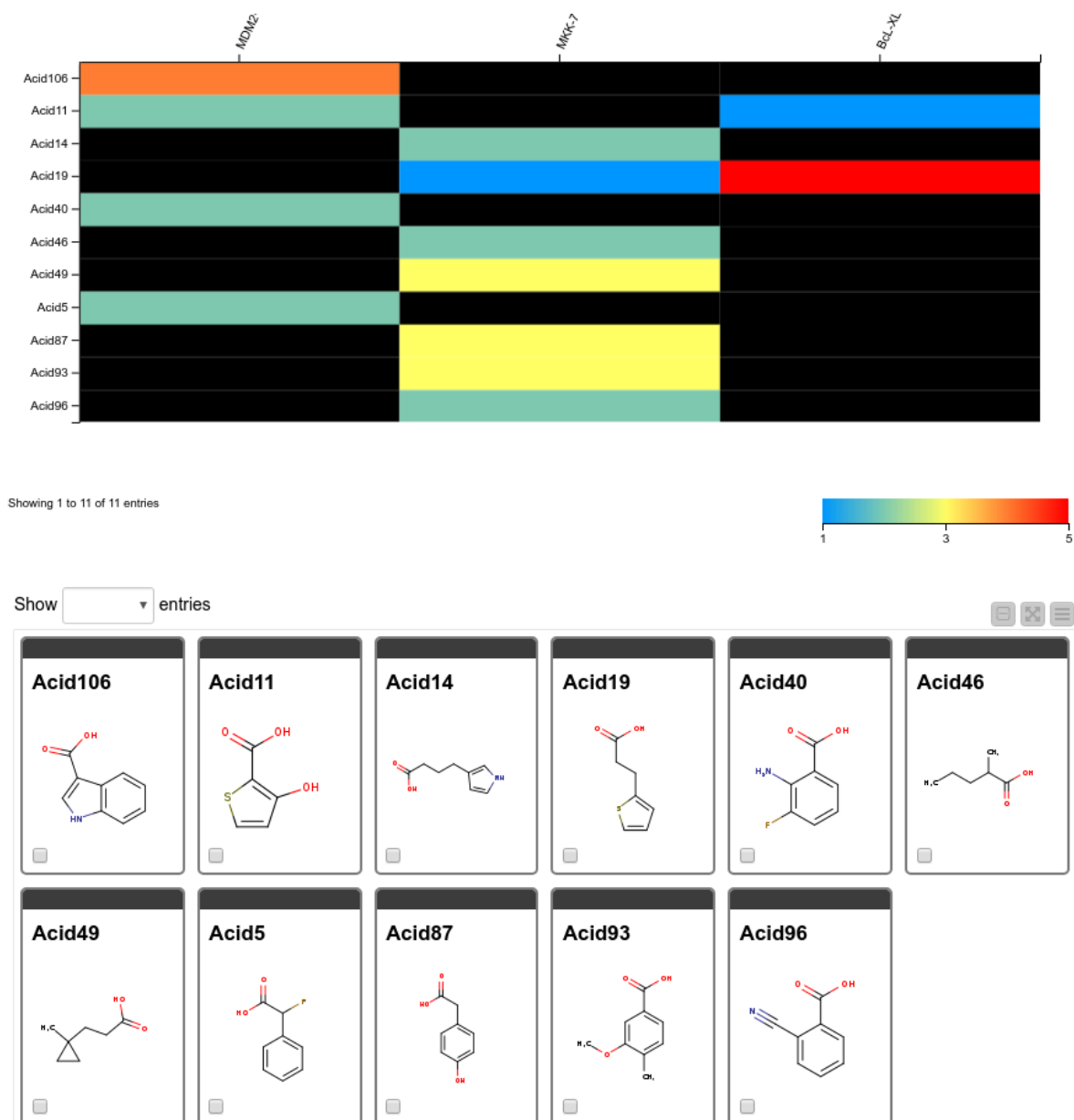


Figure 45: Correlation between the cycle 3 BBs and proteins. The cells of the heat map are color-coded according to the count of molecules presenting the specific substructure for each of the three considered proteins.

7.4.2 Docking experiments

A preliminary docking calculation was performed with SeeSAR (BiosolveIT) and then the results were verified by Maestro (Schrödinger). The two software programmes are based on different assumptions and computations, as *Hyde* is based on solvation/desolvation calculations while *Glide* is based on force fields. Therefore, molecules that showed high affinity in both simulations were considered to be more reliable and were selected for resynthesis. The selection was performed in KNIME, where the Pareto ranking algorithm was applied, so that only the molecules that performed the best in both the selection assay and the docking studies were reviewed. As for the purpose of hit resynthesis we first focused on the THQ scaffold obtained from the Povarov reaction, the results of which are reported in the following sections.

7.4.2.1 BCL-X_L

The first protein analysed was BCL-X_L whose scatter plot is shown in Figure XX. In the scatter plot each dot represents a hit from the selection assay, so data points on the top right corner describe hits with higher EF than data points on the bottom left corner. The color code corresponds to the Pareto ranking, which also considers the docking results, so molecules with high EF, high affinity, low docking score were depicted in red and were considered for resynthesis off-DNA. The molecules corresponding to the four red dots in the scatter plot are depicted in Figure 46. The two red dots on the top, indicated by the letters A and B, showed the piperazine scaffold obtained from the SnAP reaction, while the three data point with letters B, C and D were considered in this analysis as they belong to the group consisting of the THQ scaffold.

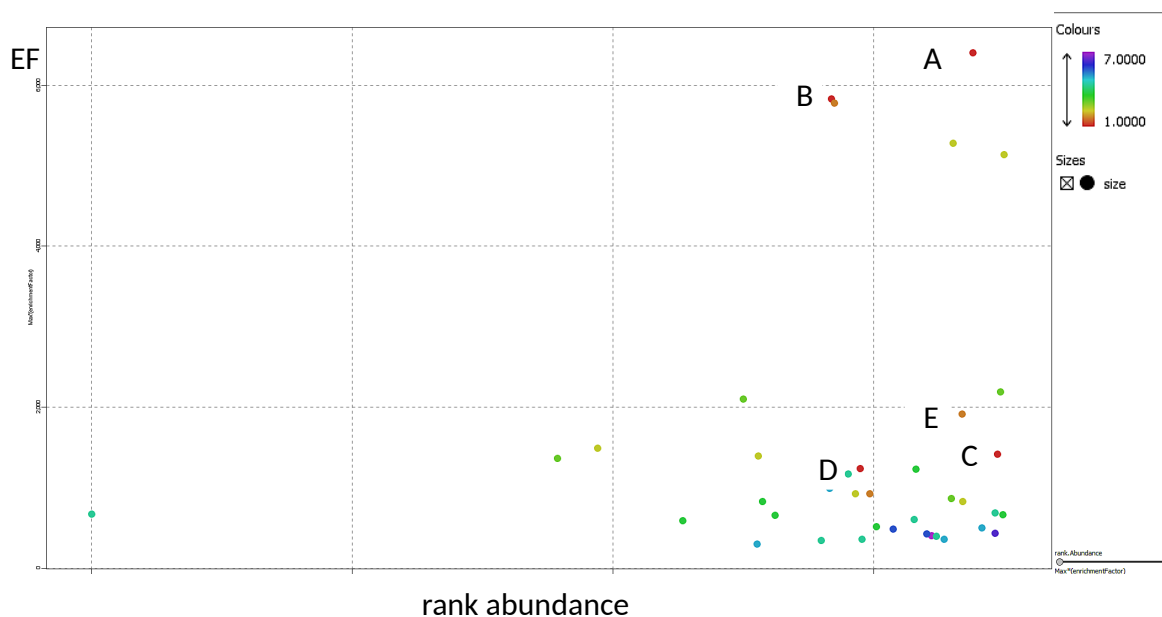


Figure 46: Scatter plot relating the EF vs RA for the protein BCL-X_L.

Although, their analysis is not covered in this thesis, it is worth noticing that the carboxylic acid inserted in the third cycle, visible in Figure 47 A and C, is common to two out of four considered hits. Although it may appear significant, the 6-fold difference in EF between the SnAP reaction-based and Povarov reaction-based *hits* is not significant because it was demonstrated in previous publications that after a certain number of cycles the PCR efficiency reaches a plateau. [155] Compounds **1**, **2** and **3** were investigated further via docking experiments.

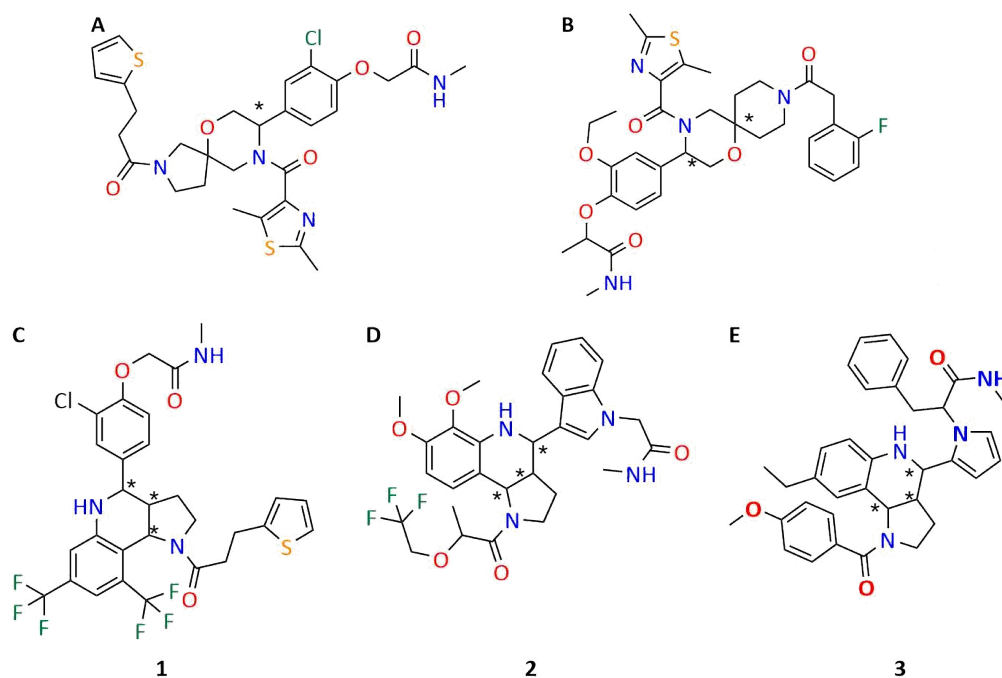


Figure 47: Hit molecules for the protein BCL-XL. (A) and (B) are based on the SnAP reaction, while (C), (D) and (E) contain the THQ scaffold obtained by Povarov reaction. The letters correspond to the scatter plot in the previous figure.

The first hit for BCL-XL with the Povarov reaction-based scaffold was compound **1** with EF of 1409. The two docking algorithms identified the same binding pocket and this could be considered already a proof of validation as this pocket was exactly the binding site of the co-crystallized ligand ABT-737. [143] Despite the slightly different perspective, in Figure 48 A and B, it is visible how the different moieties of the ligand occupied similar space within the binding pocket. In particular, the binding pocket is formed by the aminoacid residues Ala-104, Arg-139 and Glu-129 in both docking simulations, whereas the ligand differently interacts with those residues according to the two softwares. In Figure 48 A and B, the ligand is depicted after docking with *Hyde* (Figure 48 A) and *Glide* (Figure 48 B).

Hits validation by molecular docking

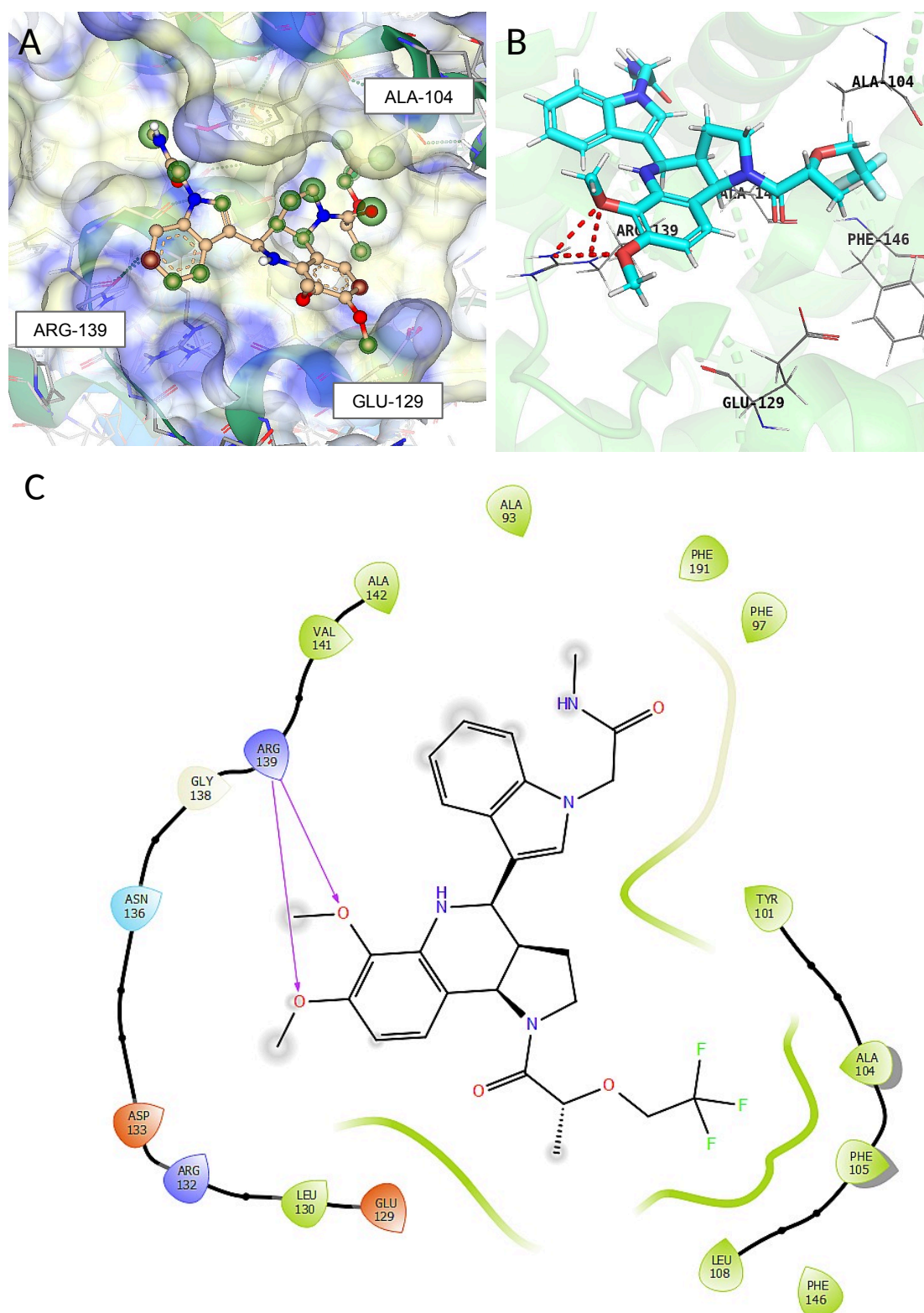


Figure 48: Docking results for compound 1. (A) Hyde 3D view of the binding site with the pose calculated by *Hyde*. To locate the binding site, important amino acids are labelled. (B) 3D view of the binding site and the pose calculated by *Glide*. (C) 2D view of the interactions calculated by *Glide*.

Overall, *Hyde* calculated an affinity falling into the nM range for this molecule. In the *SeeSAR* renderer, atoms are surrounded by an aura, coloured in green or red for favourable or unfavourable interactions, respectively. For example, the methoxy group as substituent on the aniline ring in the THQ scaffold was surrounded by a green circle due to the carboxylic acid of the Glutamate 129, which created a hydrophilic environment. According to *Hyde*, the indole moiety did not form favourable interactions most likely due to the guanidine group of Arginine 139 in close distance. In this case, a group able to accept the hydrogen bond from the protonated guanidine, such as amino or carbonyl groups, could increase the affinity. Interestingly, *Glide* did not position the above mentioned methoxy group close to Glutamate 129, but to Arginine 139 instead. The latter formed intense interactions with both hydroxyl groups on the aromatic ring. Additionally, *Glide* emphasized a hydrophobic interaction between the trifluoromethyl group of the cycle 3 acid and the residues Alanine 104, Phenylalanine 105 and Leucine 108 (Figure XX C). The three aminoacids formed a strongly hydrophobic pocket that in the original publication was occupied by the biphenyl moiety of ABT-737. [143] The *Glide* docking score was -7.1 which is a moderate result. Indeed, a score below -7 signifies that the algorithm could position the molecule with some certainty, the lower the score the higher the confidence. Importantly, in both poses the linker moiety did not occupy a deep subpocket, but only slightly protruded towards the second pocket of the binding site or was exposed to the solvent according to *Hyde* and to *Glide*, respectively. as highlighted by the grey aura surrounding the linker atoms in Figure XX C. In conclusion, despite the dissimilar interaction pattern between the ligand and the binding pocket according to the two algorithms (*Hyde* and *Glide*), the molecule still offers a similar docking pose, making it a promising candidate for resynthesis.

The second identified hit for BCL-X_L was compound **2** with EF of 1229. Similar to compound **1**, the two docking algorithms positioned it in a very similar fashion as clarified by the comparison between Figure XX A and B. *Hyde* assigned it to the μ M range of affinity and *Glide* scored it with a -8.1, meaning that this pose was more realistic than the one for compound **1**.

Interestingly, the nitrogen in position 2 of the tetrahydroquinoline ring was considered in an unfavourable position by *Hyde* (Figure 49 A), while the same atom was acting as a hydrogen bond donor towards the carbonyl group of the residue Alanine 104 when docked by *Glide* (Figure 49 B and C).

Hits validation by molecular docking

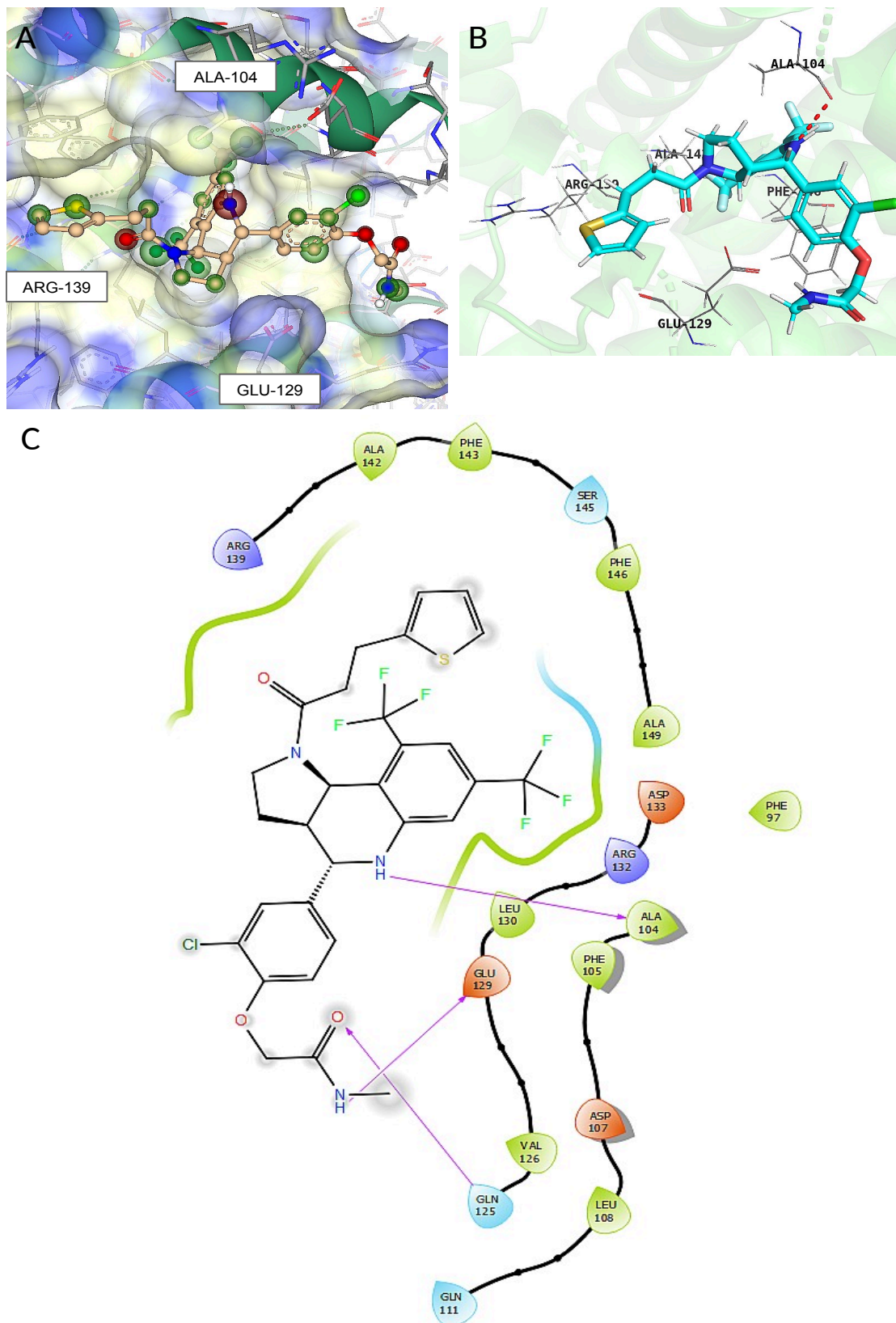


Figure 49: Docking results for compound 2. (A) Hyde 3D view of the binding site with the pose calculated by Hyde. To locate the binding site, important amino acids are labelled. (B) 3D view of the binding site and the pose calculated by Glide. (C) 2D view of the interactions calculated by Glide.

The pose assigned by *Hyde* might reflect the actual binding mode considering the oxidation process that might have taken place during library synthesis and screening. In fact, the THQ ring has a high tendency to oxidize by losing hydrogen atoms allowing it to fulfil its aromaticity. Such oxidation reactions are enabled by water as a solvent, which is predominant in the DEL context. Furthermore, the oxidation and consequent aromatization of the ring deprives this scaffold from acting as a hydrogen bond donor, while also greatly changing the polarity of the molecule. The linker, in this case, improved the interaction pattern according to *Glide*, forming a hydrogen bond with the protein backbone corresponding to Glycine 125. On the other hand, it affected negatively the affinity according to *Hyde*, highlighted by the red aura surrounding the carbonyl oxygen of the amide bond. This negative influence might be contrasted by the absence of the linker and a more hydrophobic moiety as substituent of the aromatic ring. It is worth mentioning here that the cycle 3 carboxylic acid appeared in 4 out of 18 hits. The position of the sulphur atom of the thiophene ring was well evaluated by *Hyde* as visible in the size of the aura around it and it even formed an interaction with the backbone of BCL-X_L. *Glide* did not confirmed the interaction but positioned the ring in a hydrophobic pocket defined by the residues Alanine 142 and Phenylalanine 143. Finally, the third hit for BCL-X_L isolated for the scope of this thesis was compound **3**, with an initial EF of 913, which is still in a very high range for this specific protein. In this case, the *Hyde* algorithm assigned it to the nanomolar range of affinity and *Glide* calculated a docking score of -7.2. The comparison between Figure 50 A and B exhibit the different positioning of the molecules adopted by the two docking algorithms: the cycle 3 acid and the aniline moiety which is part of the THQ scaffold are swapped in the pocket. In fact, for *Hyde* the methoxy group of the cycle 3 carboxyl acid interacted with the guanidine group of Arginine 139, while for *Glide* the same aminoacid was closer to the ethyl group of the aniline. Similar to compound **2**, the nitrogen in position 2 in the scaffold interacted with the residues in the binding pocket, in this case forming a hydrogen bond with Glutamate 129 (Figure 50 C). The linker in this case seemed to play an important role, although in both simulations resulted to be exposed to the solvent. For *Glide* the linker participated in the interaction with Glutamate 129, whereas for *Hyde* all atoms involved in the amide bond were surrounded by an evident red aura, especially the nitrogen, meaning that probably the molecule without the linker would constitute a better candidate for resynthesis and biological testing.

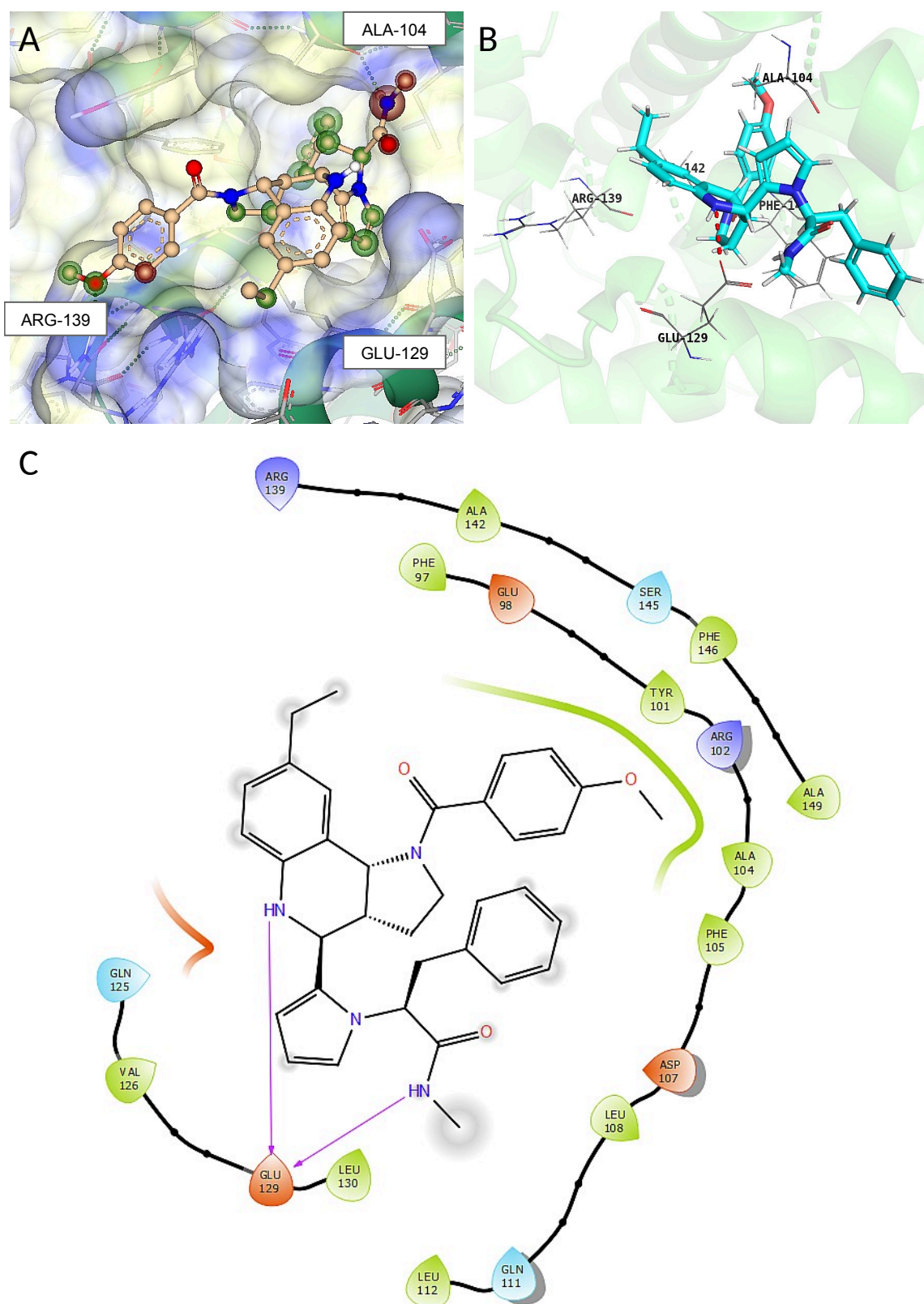


Figure 50: Docking results for compound 3. (A) 3D view of the binding site with the pose calculated by Hyde. To locate the binding site, important amino acids are labelled. (B) 3D view of the binding site and the pose calculated by Glide. (C) 2D view of the interactions calculated by Glide.

7.4.2.2 MKK7

The analysis of the sequenced hits for the target MKK7 produced the scatter plot in Figure 51 displaying the EF vs the rank abundance as axes and the hit molecules as dots coloured according to the Pareto ranking. In particular, red dots refer to molecules with high EF, affinity in the nM or μ M range and low docking score. The molecules corresponding to the red dots highlighted with the letters in the plot are depicted in Figure 52 under the corresponding letter. Molecules in Figure 52 A, B and D presented SnAP reaction-based scaffolds so they did not belong to the scope of this thesis, while the one in Figure 52 C presented the THQ scaffold, so it was further analysed as compound 4.

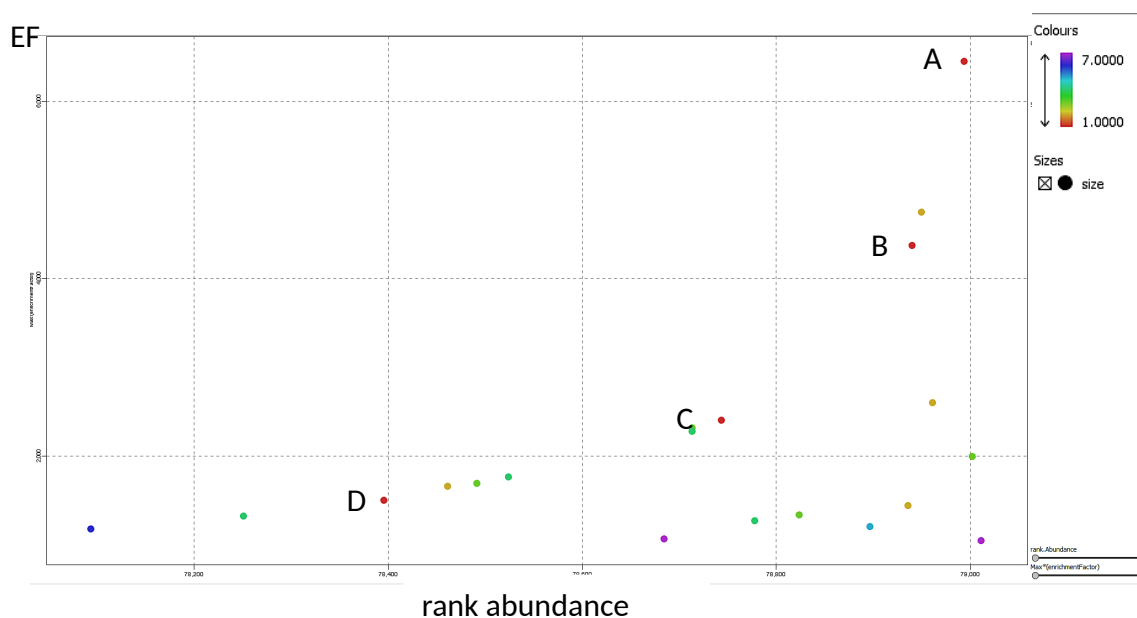


Figure 51: Scatter plot relating the EF vs RA for the protein MKK7.

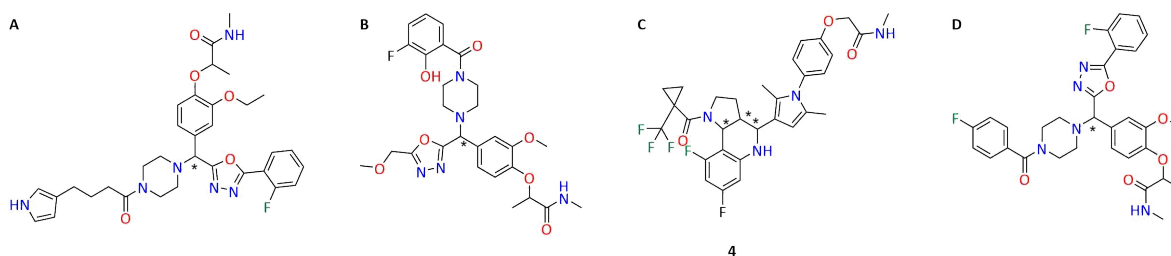


Figure 52: Hit molecules for the protein MKK7. (A), (B) and (D) are based on the SnAP reaction, while (C) contains the THQ scaffold obtained by Povarov reaction. The letters correspond to the scatter plot in the previous figure.

It performed well in the selection assay, with an EF of 2393, and was excellently evaluated by both docking algorithms.

Hyde assigned it a nanomolar range of affinity and *Glide* a docking score of -10.1, which underlined high confidence in the positioning of the molecule.

For this hit, both algorithms agreed on the interaction pattern, involving the cycle 1 aldehyde moiety with two residues in the binding site. In details, *Glide* highlighted a pi-cation interaction between the pyrrole ring and Lysine 165. The Methionine 215, on the other hand, formed hydrogen bonds with the nitrogen of the linker moiety according to *Hyde* (Figure 53A) and with the oxygen in *para* position to the pyrrole on the benzyl ring according to *Glide*, as emphasized by the yellow arrows in Figure 53A and C respectively.

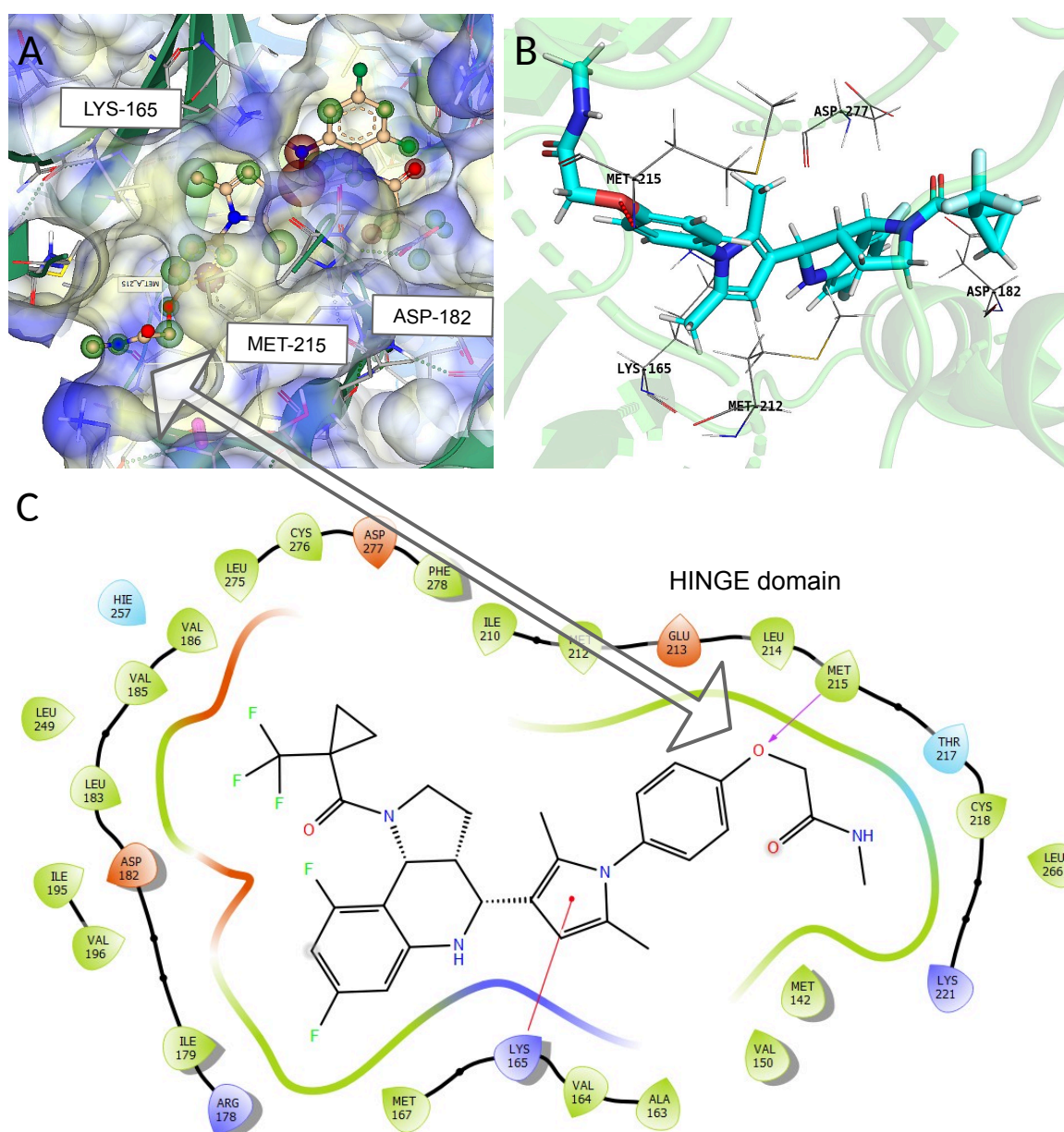


Figure 53: Docking results for compound 4. (A) 3D view of the binding site with the pose calculated by Hyde. To locate the binding site, important amino acids are labelled. (B) 3D view of the binding site and the pose calculated by Glide. (C) 2D view of the interactions calculated by Glide.

Interestingly, the linker moiety points towards the Cysteine 218, used for the covalent bond for inhibitors such as Ibrutinib. Therefore, in the case of a covalent inhibitor, the acrylamide could be reacted with the carboxylic acid of the cycle 1 building block to obtain the same effect. The fact that two docking algorithms based on so different assumptions predicted a very similar pose with high confidence is a clear sign that compound 4 was worth being tested in biological assays.

7.4.2.3 MDM2

The calculation and comparison of the EFs for the third protein MDM2 produced the scatter plot in Figure 54. It is noticeable the fourfold gap in EF between the first and the second highest enriched compounds, which might lead to consider the first molecule an outlier. Although this hit consisted of a SnAP reaction-based scaffold, a closer look elucidated that the cycle 3 acid contained a pyrazolo[1,5- α]pyridine moiety (Figure 55 A) which showed high structural similarity with the indole group. This group has been considered an anchor for this target as its respective aminoacid, Tryptophan, is involved in the interaction between MDM2 and p53. This finding supported the validation of the selection assay because the indole-based carboxylic acid in the third cycle of the library synthesis accounted for 4 out of 21 identified hits for MDM2 (Figure 55). Since the software SeeSAR could not assign any of the Povarov reaction-based hits to the nM or μ M range, two hits with high EF and good *Glide* docking score were reported in Figure 56 with the respective interaction patterns.

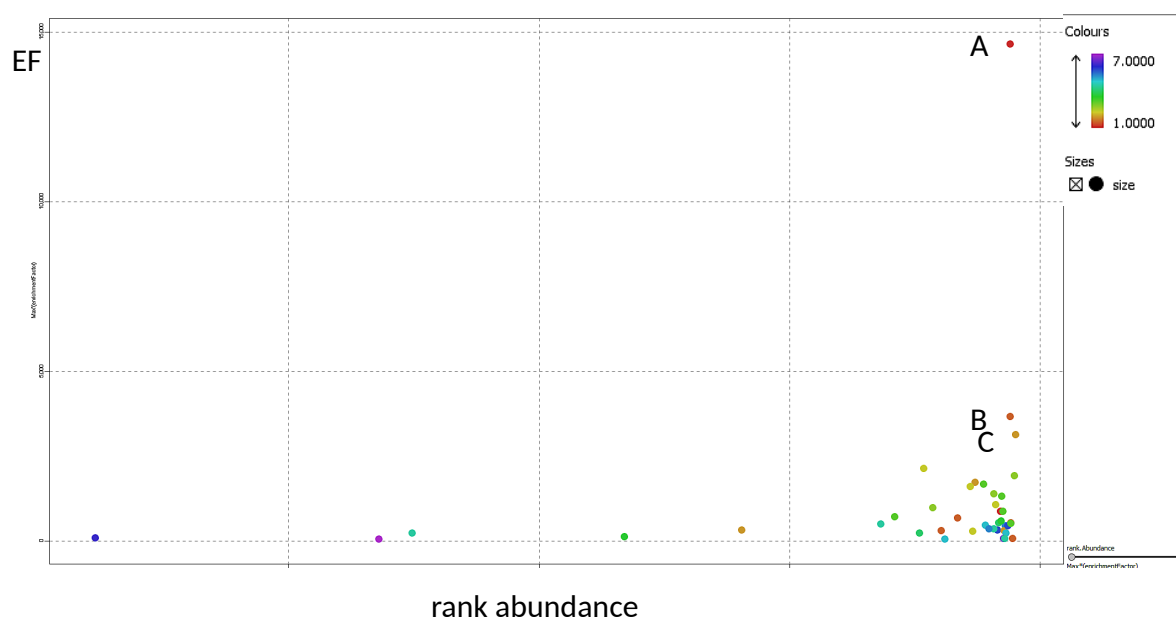


Figure 54: Scatter plot relating the EF vs RA for the protein MDM2.

Hits validation by molecular docking

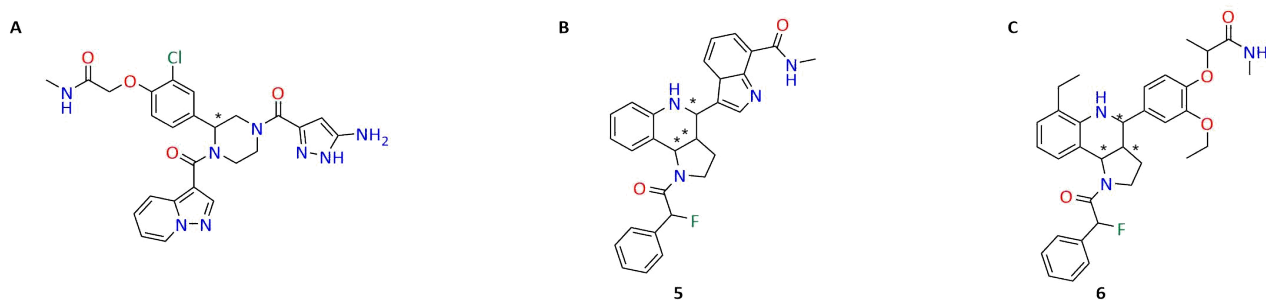


Figure 55: Hit molecules for the protein MDM2. (A) is based on the SnAP reaction, while (B) and (C) contain the THQ scaffold obtained by Povarov reaction. The letters correspond to the scatter plot in the previous figure.

Compound **5** (Figure 56 A and C) contained the indole group at the aldehyde for the first cycle of library synthesis, which fit deeply inside the binding site. As the linker moiety was always bound to the first cycle aldehyde, it was also buried inside the pocket, rendering the pose unrealistic in the context of the selection assay. However, the docking score of -9.2 might suggest that a synthesis omitting the linker could lead to successful biological experiments. In compound **6** (Figure 56 B and D), the linker is exposed to the solvent and the docking score was -8.9.

Hits validation by molecular docking

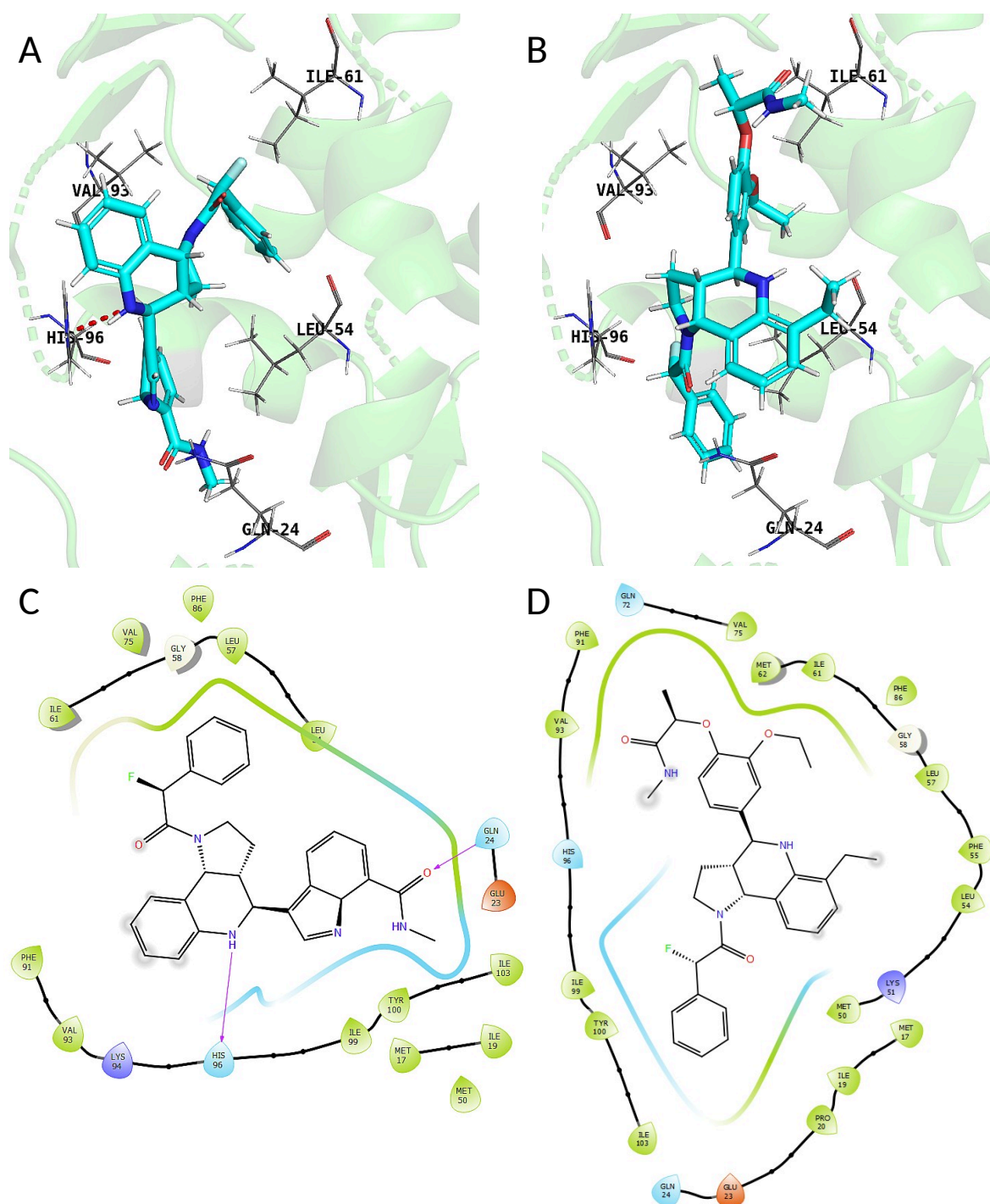


Figure 56: Analysis of the docking results for compound 5 and 6. (A) and (B) Glide 3D view of the molecules in the binding pocket and (C) and (D) 2D interaction schemes.

The two hits display the same carboxylic acid, the α -fluorobenzenacetic acid, which occupied different spaces in the pocket. Finally, the two compounds constituted good candidates for resynthesis firstly for their high EFs, 1701 and 1578 respectively, and in addition for their good docking scores.

The results of the whole analysis are summarized in Table 21.

Table 21: Results of the whole analysis.

Protein	Compound	EF	SeeSAR nM affinity range	Glide docking score
	1	1409	yes	-7.1
BCL-XL	2	1229	no	-8.1
	3	913	yes	-7.2
MKK7	4	2393	yes	-10.1
	5	1701	no	-9.2
MDM2	6	1578	no	-8.9

7.5 Conclusions

In this chapter, chemoinformatics and computational methods have been employed to interpret the calculations deriving from the selection assay of a DEL screening campaign. In the process, supplementary analyses were performed. At first, frequent hitters were identified as molecules binding more than four proteins and considered as false positives. Further investigation on these molecules might reveal interactions with the material used in the selection assay. In particular, the substructure *1,3-dihydro-2H-imidazol-2-one* was found in all *frequent hitters*, highlighting its high reactivity. In addition, trends could be detected in terms of preference from the proteins side towards specific substructures, such as the known MDM2-indole predilection. Once the hit molecules were refined, they were subjected to the docking protocol, with two aims. Firstly, the docking protocol was intended as validation method for the selection assay. In fact, only molecules performing well in all practical and simulated experiments were selected for resynthesis and further biological test. Two docking algorithms were employed, which are based on very distinct assumptions, rendering the validation protocol more robust. Additionally, in case a hit resulted active against a target, the docking algorithm could suggest the binding mode, which would otherwise be unknown.

The procedure was recurred for three proteins and focused on the molecules showing the THQ scaffold obtained with the Povarov reaction. While for the protein BCL-X_L the docking algorithms and the selection assay agreed on three molecules, for the protein MDM2 the first docking algorithm could not identify any affine molecules in contrast with the second docking procedure which showed encouraging results. For the last considered protein, MKK7, the selection assay and the docking protocols strongly agreed on one molecules from the THQ scaffold class, which was definitely worth to be further analysed.

Finally, not only the single *hits* are going to be submitted to further testing, but also series of congeneric compounds carrying features that could improve the affinity, based on the docking results. The selected molecules are currently under synthetic procedure and will be analysed in bioactivity assays.

8 *General conclusions and future perspectives*

Over the course of this thesis, chemoinformatics and computational methodologies have been employed to support and optimize the technology of DNA-encoded libraries from the design to the validation. In particular, the chemoinformatics aspects of such objective were accomplished within the KNIME Analytics Platform. In the first place, a novel algorithm was designed to select potential chemistries from large databases and employ them in the DEL synthesis. One chemistry that was extracted by the algorithm could be successfully applied to DNA-encoded substrates and it is currently under investigation as candidate for the synthesis of a DNA-encoded library. In parallel, a workflow for selecting building blocks was generated as well. The main aim of such selection process was the products' chemical and geometrical diversity and the building blocks were adapted to chemical transformations that have been already established for DELs. Finally, the respective libraries were enumerated and compared with drug-like compounds to assess their significance in a drug discovery project. The combination of multiple chemistries and diverse building blocks guaranteed a wide coverage of chemical space, which is a promising feature in terms of hits identification. Lastly, the validation of DEL hits was aided by docking, for the purpose of reducing noise and removing false positives. At first, the sequencing results were analysed with KNIME and the data set could be decluttered from *frequent hitters*. Hence, the remaining structures were docked onto the target proteins to assess their affinity and their binding modes. The chemoinformatics and docking results guided towards identifications of trends and optimization of the molecules. As mentioned in the introduction, the results of the chemoinformatics analysis shall guide scientists towards data-driven solutions and optimized design of experiments, as highlighted by the green arrow in Figure 57. In Figure 57, the main contributions to each stage of the DEL process are highlighted on top. The design of the library can certainly be aided by machines and databases, whereas the synthesis and screening requires manual effort, although progress in the direction of lab automation has been made. [156], [157], [158] The validation of hits, beside biological tests, can be performed *in silico* as well via docking or via machine learning methods.

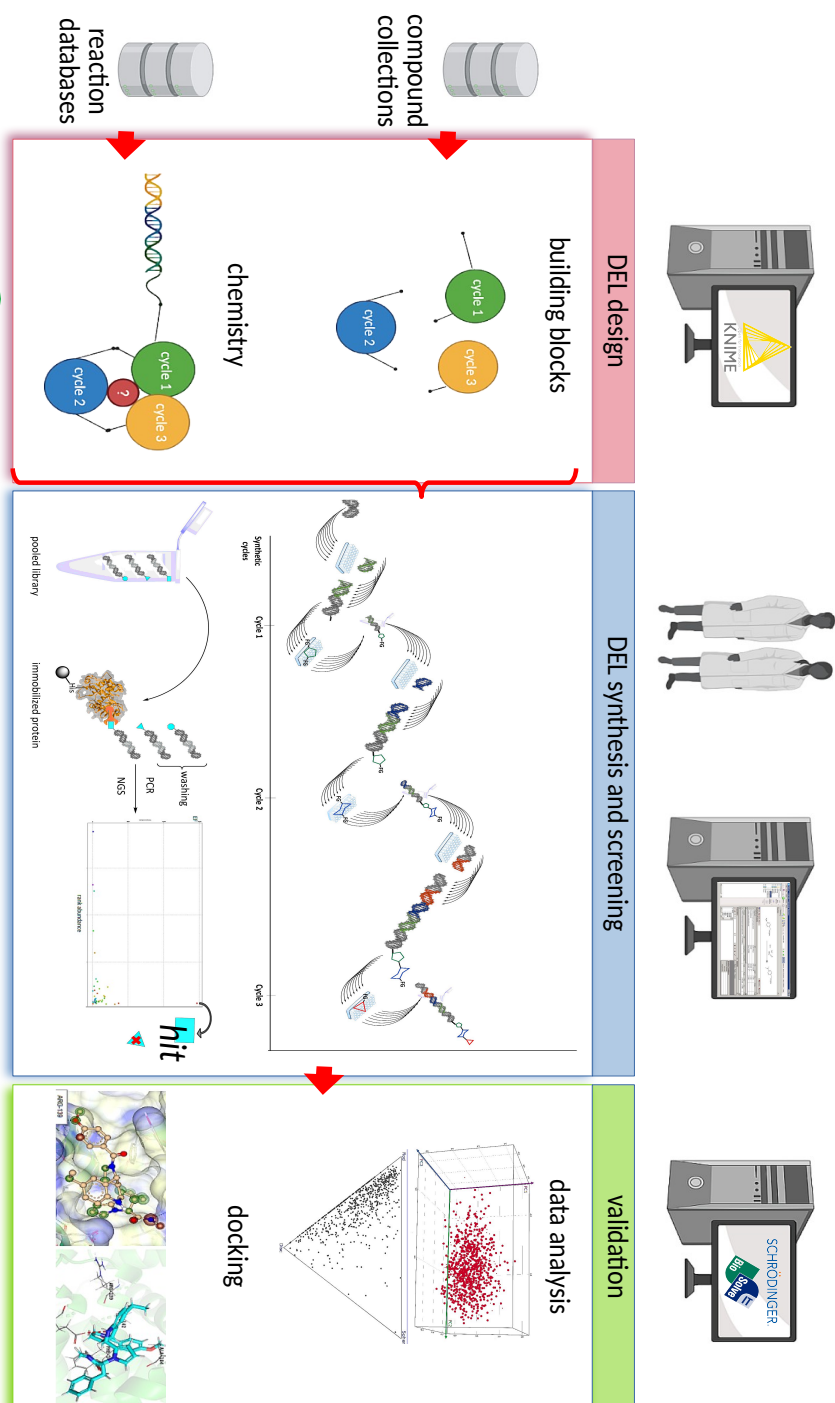


Figure 57: Whole computer-aided DEL process. The DEL design has been supported by KNIME for the selection of the chemistry and the building blocks from chemical databases. On the other hand, the synthesis and selection of the library has not been supported by machines so far, whereas the identified hits have been validated by docking experiments, coupled with ongoing biochemical tests. Inside the monitors, the software utilized for the considered step is specified, except for the screen in the centre, where instead there is a screenshot of the Fujitsu Electronic Laboratory Notebook System. This represents a guideline that would improve consistency in the DEL process. The insights gained from the chemoinformatics analysis and computational simulations can serve the data-driven design of further DEL design (green arrow). Indeed, building blocks that caused unspecific interactions will be avoided in the future as well as chemistries which do not guarantee a large coverage of chemical space. Part of the figure has been generated by Biorender.com.

However, two important notes are to be made regarding the computer-aided design and synthesis: databases needs large efforts in data curation. Nowadays, chemistry databases are largely based on mining scientific articles, yet laboratory journals constitute a key source of primary data, including negative results that are so precious to machine learning methodologies. [159], [160], [161] In fact, suboptimal results are as fundamental as successful results for models to make reliable predictions. For example, intense efforts are directed towards the creation of a freely available data set of balanced reactions. [162] The text mining technologies would be enhanced by routinely using a universal, unambiguous and machine-readable chemical language. [163] Such system would improve reproducibility not only by humans but also by robots, lifting the chemists from the burden of repetitive or dangerous procedures. This concept has been already applied to flow-chemistry [164], but it is currently investigated for a broader scope of reactions and approaches. [165]

Lastly, the unification of all stages of the DEL technology under one workflow would be possible if one common platform was utilized for the purpose, together with a universal chemical language which could be decoded by both humans and machines. The common platform could be represented by KNIME, for its simplicity and all the extensions that are currently implemented, including biology tasks, docking and machine learning. In summary, a computer-aided DEL process has been reported, which can still be improved by unifying all modules under one workflow and eventually coupling it with laboratory automation. However, for this purpose the adjoining infrastructure, such as electronic laboratory notebooks, is to be implemented as well.

9 Experimental part

9.1 Chemistry selection

9.1.1 KNIME workflow and tables

The complete KNIME workflow with a sample of 100 reactions from the Reaxys® database is available at the following link: <https://kni.me/s/R8gFmu9rgDDqVxtp>.

In the Appendix at the end of the thesis, the tables utilized over the course of the workflow are listed as KNIME reports, generated with the BIRT extension.

9.1.2 Procedures for the selected reactions and analytical data

9.1.2.1 Materials and instruments

Unless otherwise noted, chemicals were purchased from *abcr*, *Acros Organics*, *Alfa Aesar*, *Fisher Scientific*, *Merck*, *Sigma Aldrich*, *TCI* and *VWR* and were used as provided without further purifications. Dry solvents (MeCN, DCE, DMA, DMF, DMSO, EtOH) were used as commercially available.

5'-Aminolinker-modified DNA oligonucleotides on controlled pore glass solid support (CPG, 1000 Å porosity) were synthesized by *Ella Biotech GmbH* (Planegg, Germany).

CPG with oligonucleotide-small molecule conjugates were filtered and washed through synthesis columns using a vacuum manifold (Vac-Man®) from *Sigma Aldrich*.

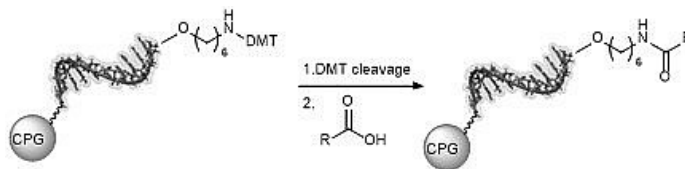
Analytical RP-HPLC: HPLC analysis was performed on a *Shimadzu Prominence* using a C₁₈ stationary phase column (Phenomenex, Gemini; 4.6 x 100 mm, 110Å, 5 µm) and a gradient of 10mM aqueous triethylammonium acetate (pH=8.0)/MeOH. HPLC traces were recorded at 254nm wavelength.

Method: Step gradient of 10% to 60% MeOH in 10mM aqueous triethylammonium acetate (pH=8.0) within 22 min at a flow rate of 0.6 mL/min.

MALDI-TOF: Mass analysis was performed on a MALDI TOF/TOF MS from *Bruker Daltonics* using 2',4',6'-trihydroxyacetophenone (THAP) matrix (*Dichrom*).

9.1.2.2 General procedures (GP)

9.1.2.2.1 Amide coupling (GP1)

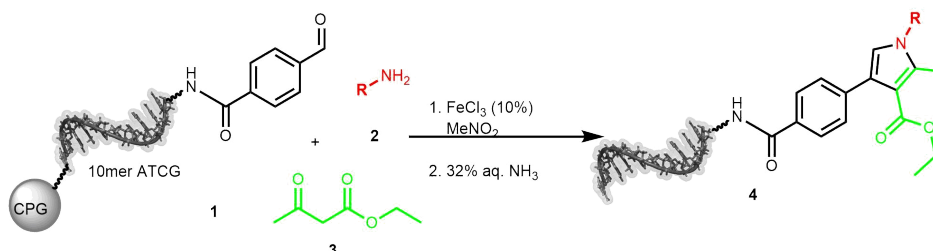


Step 1: DMT-protecting group of CPG-bound oligonucleotide (250 nmol, 9-10 mg of solid phase material) was removed by addition of 200 μL 3% trichloroacetic acid in DCM for 1 min. An orange colouring of the solution indicated successful removal of protecting group. The deprotection was repeated 3-5 times until no further colouring of the solution was observed. CPG-bound deprotected DNA was washed three times with each 200 μL of 1% TEA in ACN, DMF, MeOH, ACN and DCM and dried in vacuo.

Step 2: CPG-coupled oligonucleotide, carboxylic acid and HATU were dried in vacuo for 30 min. Stock solutions of all reactants in dry DMF were prepared before the reaction was started. HATU (25 μmol , 100 equiv., 111 mM calculated for the final volume of 225 μL) dissolved in 75 μL dry DMF and DIPEA (62.5 μmol , 250 equiv., 277 mM calculated for the final volume of 225 μL) were added to the solution of carboxylic acid (25 μmol , 100 equiv., 111 mM calculated for the final volume of 225 μL) in 75 μL dry DMF. The mixture was shaken for 5 min and added to CPG-coupled DNA (250 nmol, 1 equiv.) suspended in 75 μL dry DMF. The amide coupling reaction was shaken at ambient temperature for 2 h. Next, the CPG-coupled conjugate was filtered over a filter column, washed three times with each 200 μL of DMF, MeOH, ACN and DCM and dried in vacuo. Amide coupling was repeated two times. Completeness of amide coupling was controlled by cleaving off a small portion of CPG-coupled oligonucleotide conjugate (0.7–0.9 mg, \sim 20 nmol) with 500 μL AMA (AMA = aqueous ammonia (30%) / aqueous methylamine (40%), 1:1, vol/vol) for 30 min (TC) or 4 h (ATGC-sequences) at ambient temperature. Afterwards 20 μL of 1 M Tris buffer (pH = 7.5) were added, the mixture was dried under reduced pressure (SpeedVac) and DNA was dissolved in 200 μL distilled water. The crude reaction mixture was analysed by analytical RP-HPLC and MALDI-MS. In case of uncompleted coupling (< 90%) the reaction was repeated a third time. Unreacted amines were capped with acetic acid anhydride (three times 200 μL , 30 s, 1:1 mixture of THF/methylimidazole, 9:1, vol/vol, and THF/pyridine/acetic

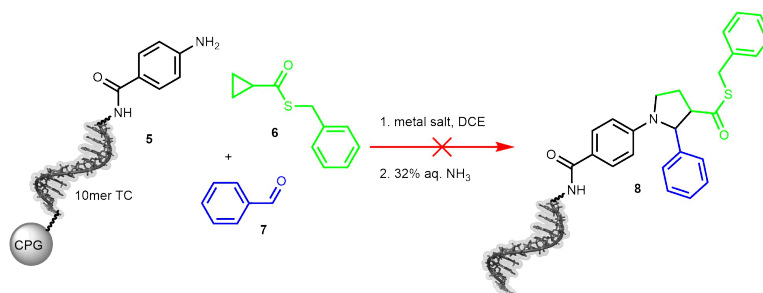
acid anhydride 8:1:1, vol/vol). The capped CPG-coupled oligonucleotide conjugate was washed three times with each 200 μL of DMF, MeOH, ACN and DCM and dried in vacuo.

9.1.2.2.2 Pyrrole synthesis on DNA-aldehyde conjugate (GP2)



The amine **2** (200 eq., 4 μmol), the ethylacetoacetate **3** (200 eq., 4 μmol), and FeCl_3 (10%, 20 eq., 0.4 μmol) were solved in nitromethane and added to 20 nmol of the CPG-bound DNA for a final volume of 40 μL and a concentration of 250 μM for all reagents except the metal salt whose concentration was 25 μM . The reaction mixtures were stirred, then transferred to a filter column and washed with dimethylformamide. Firstly, the samples were incubated with EDTA for 30 seconds and then washed 3 times with 200 μL dimethylformamide, metanol, acetonitrile and dichloromethane. Finally, the DNA-conjugates were cleaved from the solid support by dispersion in 32% aq. solution of ammonia, which was stirred for four hours. Subsequently, 20 μL of 1 M Tris buffer (pH = 7.5) were added and the mixture was dried under reduced pressure (SpeedVac). After re-dispersion in ddH₂O and filtration, the samples were further purified by adding 5 μL of 1,3,5-Triazine-2,4,6-trithiol trisodium salt solution to 45 μL of sample and stirring for 30 minutes. Then, the dispersion was centrifuged for 30 min at 4 $^\circ\text{C}$ at 11000 rpm and to the supernatant 5 μL of 3% sodium acetate aq. solution and 220 μL ethanol were added. The solution was then incubated at -80 $^\circ\text{C}$ overnight, then centrifuged, decanted and the pellet was re-dispersed in 200 μL of ethanol, incubated for 1 hour at -80 $^\circ\text{C}$ and then centrifuged and decanted. The pellet was solved in water and analysed by RP-HPLC and MALDI-TOF.

9.1.2.2.3 Pyrrolidine synthesis on DNA-amine conjugate (GP3)

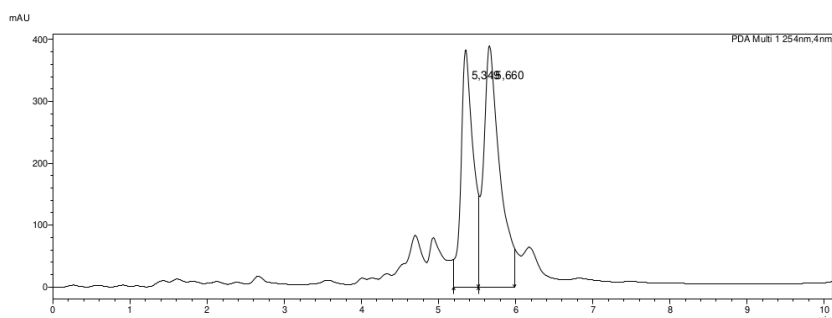


The oligonucleotide-amine conjugate on CPG (20 nmol), the thioester **6** (1000 eq., 20 μ mol) and the benzaldehyde **7** (1000 eq., 20 μ mol) were dried under low pressure for three hours, then they were separately solved in dichloroethane. The reagents including the metal salt (1000 eq., 20 μ mol) were added to the DNA/dichloroethane dispersion, and the reaction was stirred. Afterwards, the reaction was stopped by washing the CPG-bound DNA three times with 200 μ L dimethylformamide, methanol, acetonitrile and dichloromethane. For the analysis, the DNA conjugate was cleaved from the CPG by dispersion in a 32% aq. solution of ammonia and incubation for 30 minutes. Then, 20 μ L of 1 M Tris buffer (pH = 7.5) were added, the mixture was dried under reduced pressure (SpeedVac) and redissolved in water to filter the CPG. The water phase was analysed by MALDI-TOF to assess the presence of product.

9.1.2.3 Analytical data

DNA conjugate 1: CPG-bound 10mer ATCG-(CH₂)₆-NH₂ conjugate was reacted with 4-formyl-benzoic acid according to the procedure GP1.

HPLC trace of crude reaction mixture **7** (Analytical RP-HPLC).



MALDI-MS spectrum of crude reaction mixture **1**

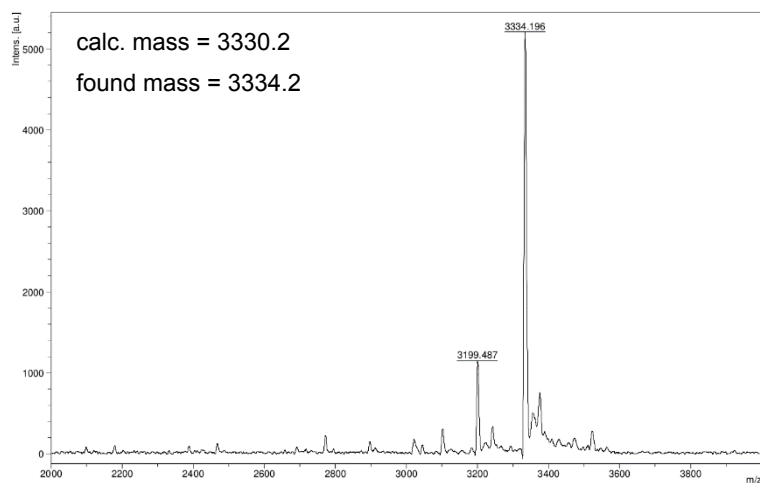
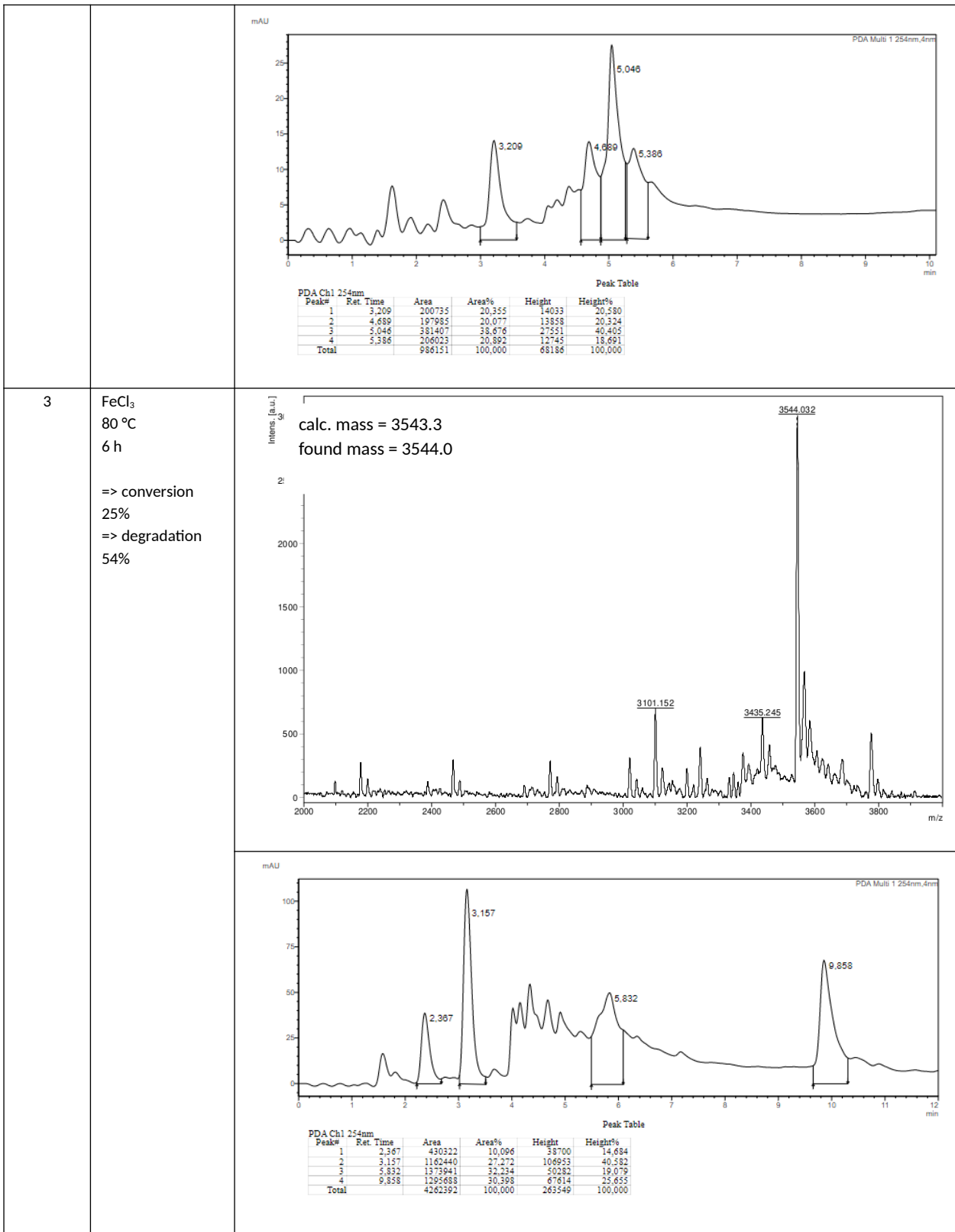


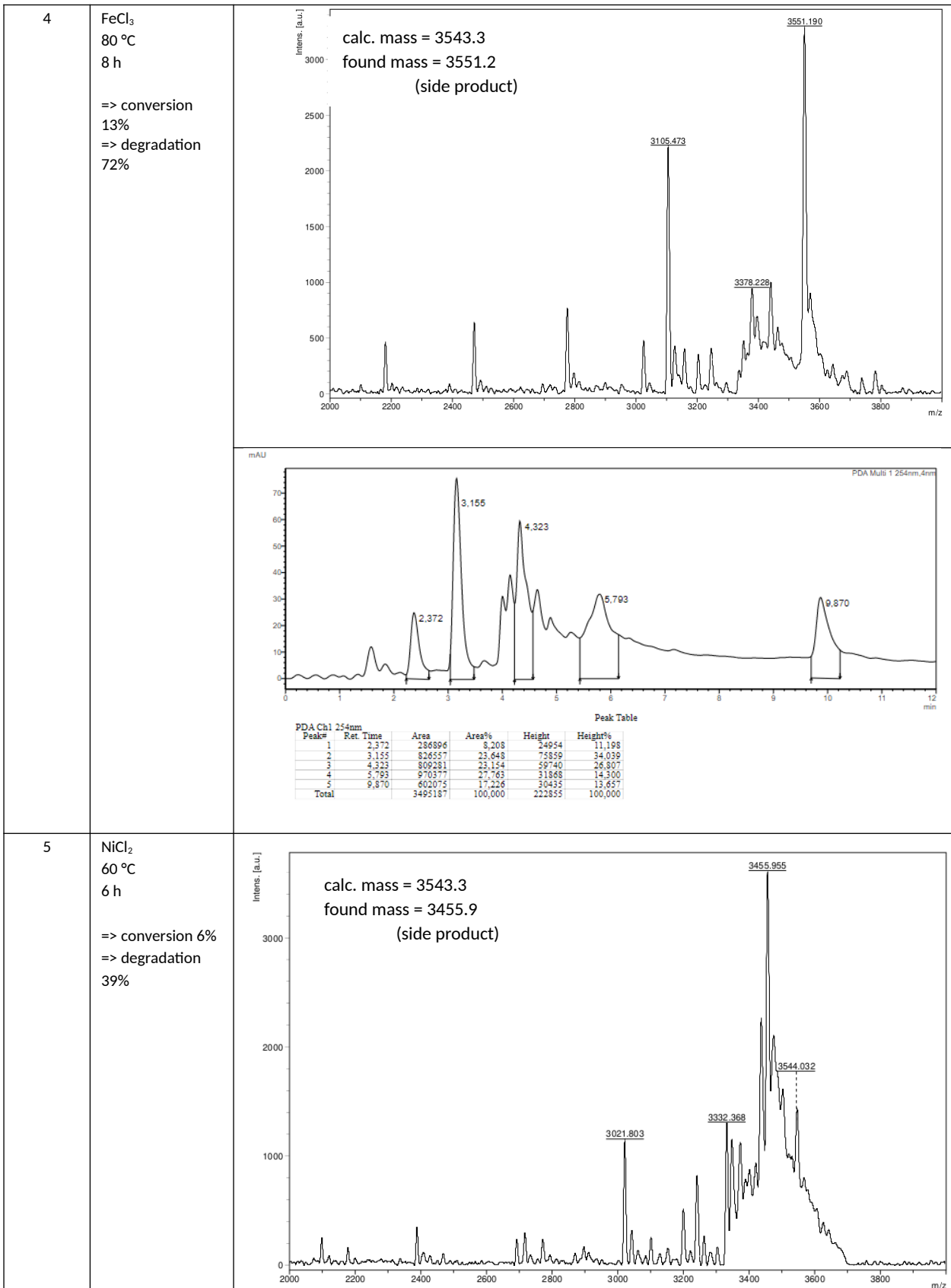
Table 22: Optimization of pyrrole synthesis on a CPG-bound DNA-aldehyde conjugate **1** according to **GP2**.^a

Entry	Reaction conditions	MALDI spectrum of the crude reaction mixture RP-HPLC trace of the crude reaction mixture
1	<p>FeCl₃ 60 °C 6 h</p> <p>=> conversion 0% => degradation 20%</p>	

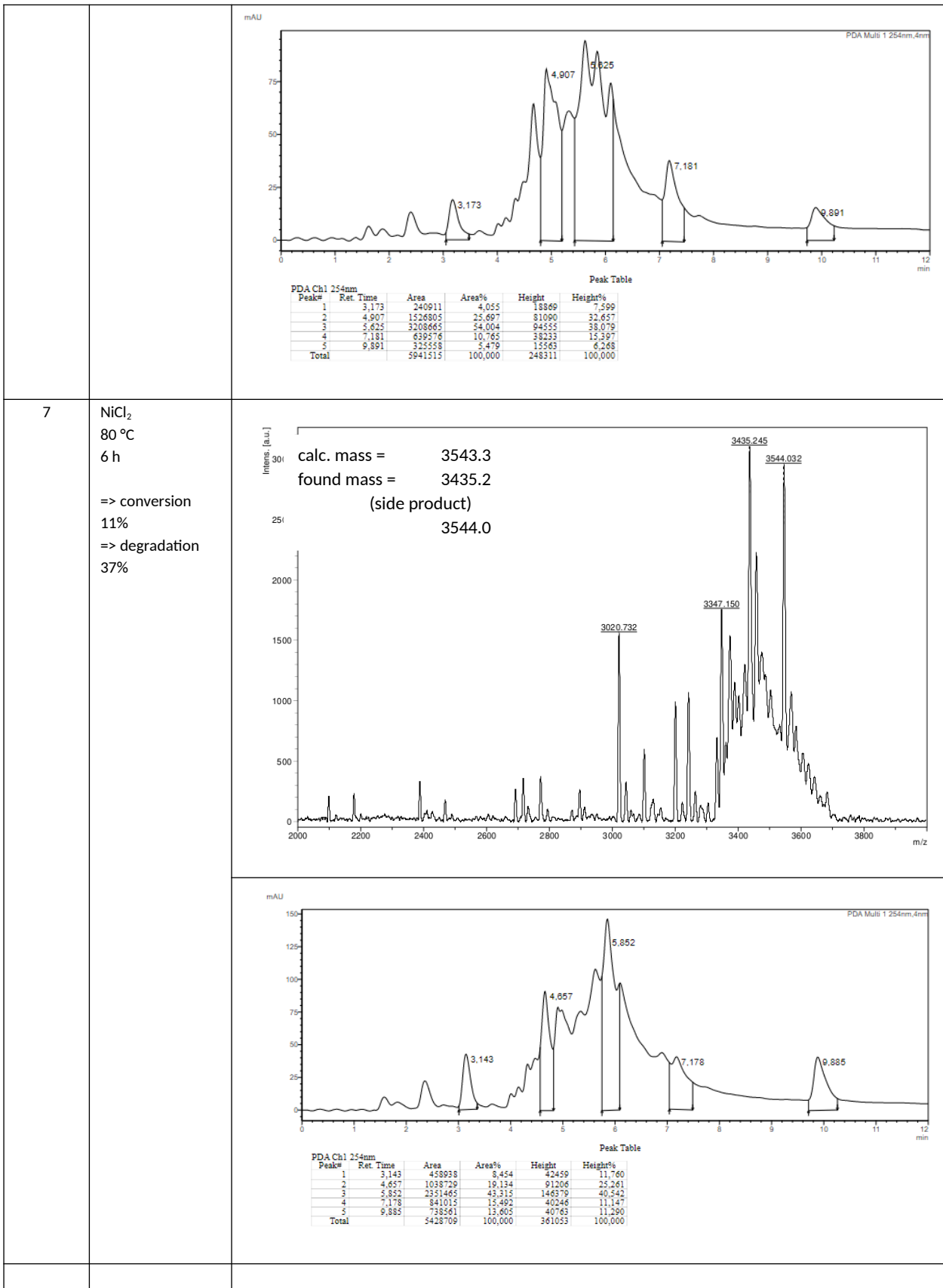
Experimental part



Experimental part



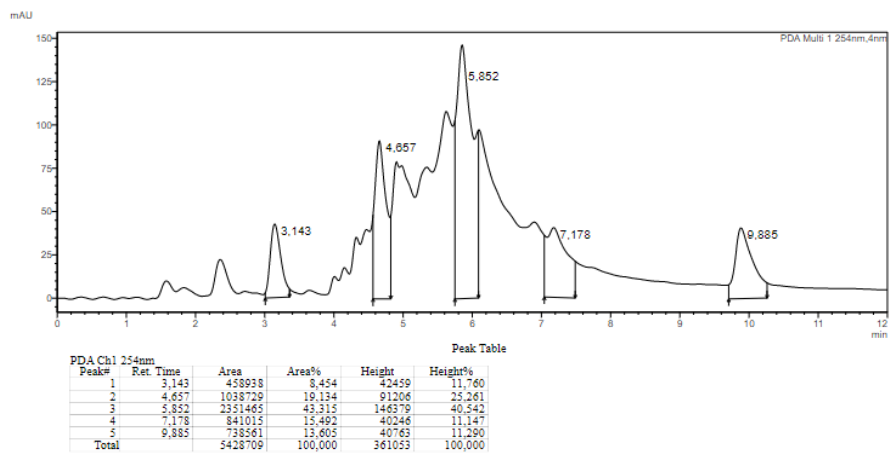
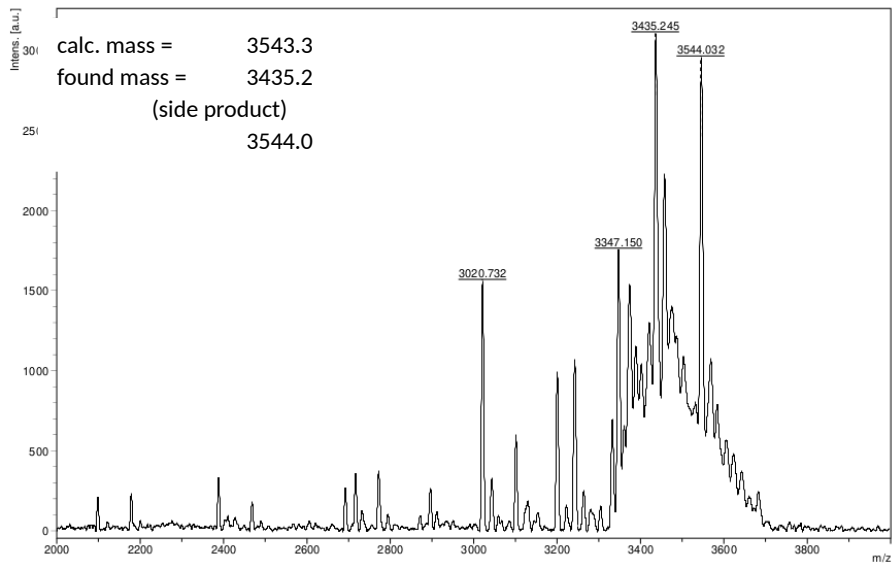
Experimental part



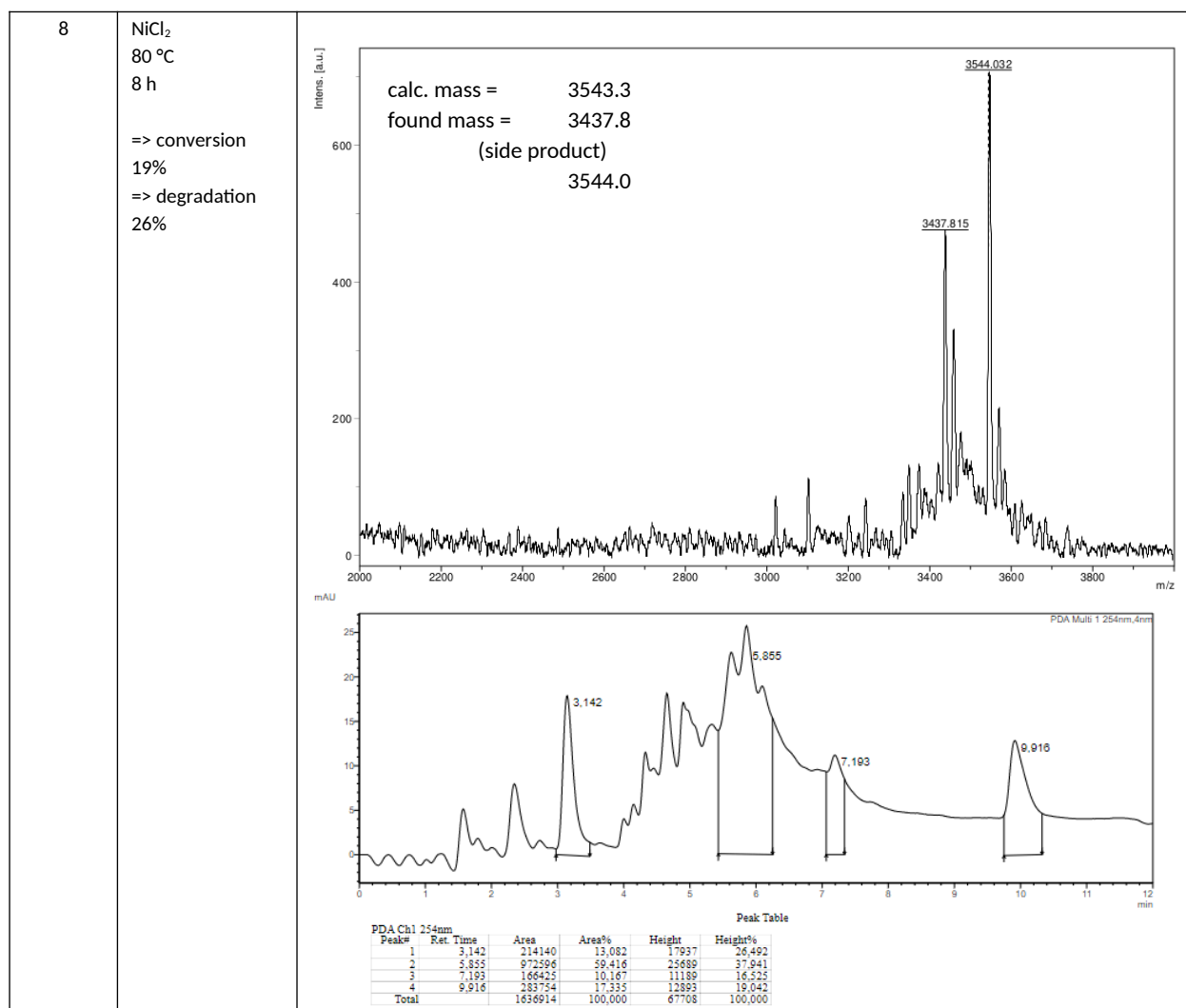
7

NiCl₂
80 °C
6 h

=> conversion
11%
=> degradation
37%



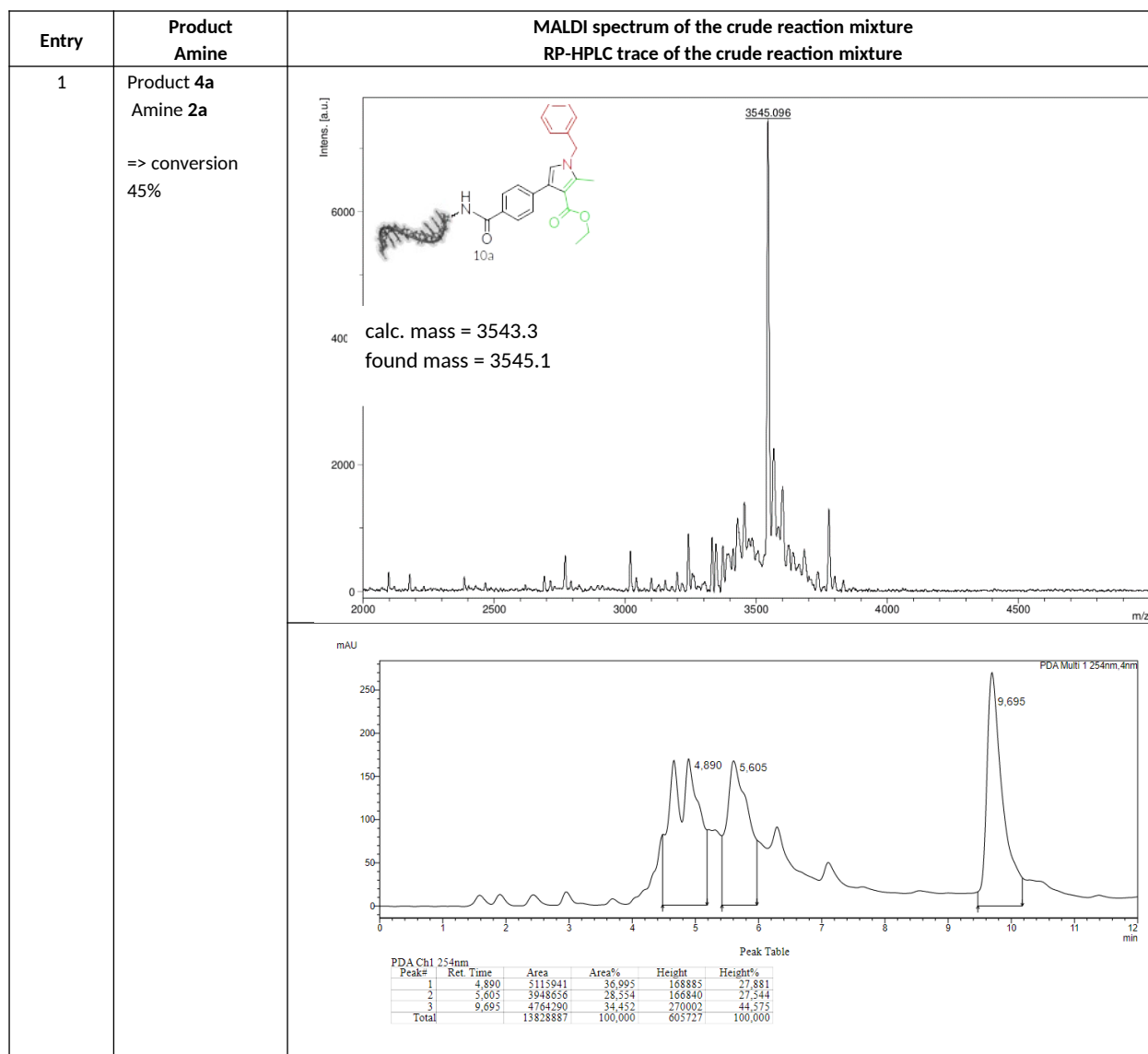
Experimental part



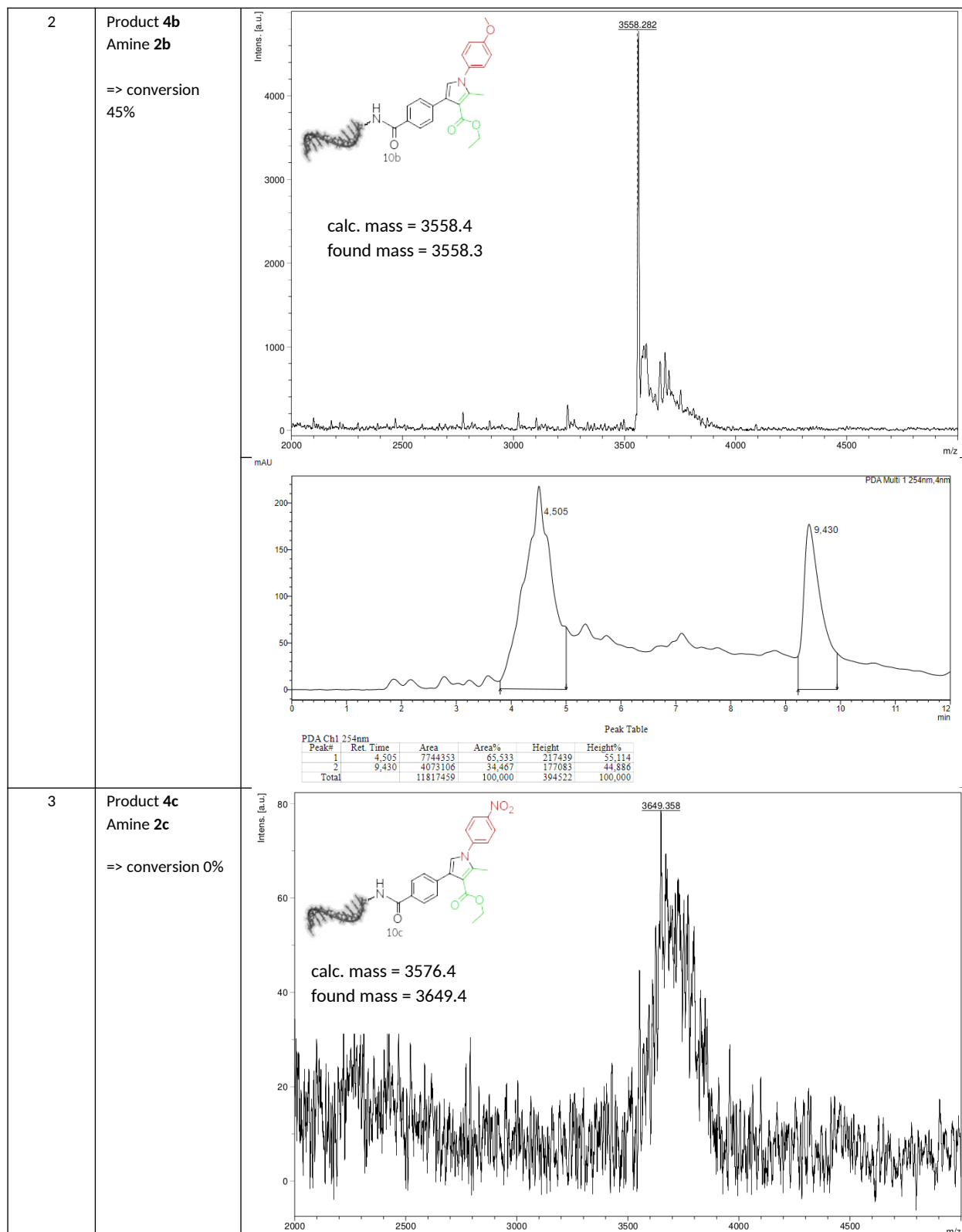
^a The suspension of DNA-aldehyde conjugate **1** (20 nmol), benzylamine **2a** (250 μM), ethylcetoacetate **3** (250 μM) and metal salt (25 μM) in solvent (40 μL) was shaken at the given temperature for the given time. DNA cleavage with 32% aq. Ammonia solution at ambient temperature for 4 h. ^b Determined by RP-HPLC analysis based on the ratios of **1** to **4a** (AUC). ^c Determined by RP-HPLC analysis and scaled to the purity of the CPG-coupled DNA-aldehyde conjugate **1**. MeNO₂ = nitromethane. N.d. = not detected, 10mer ATGC = 5'-NH₂-C₆-GTCATGATCT-3'-CPG.

Experimental part

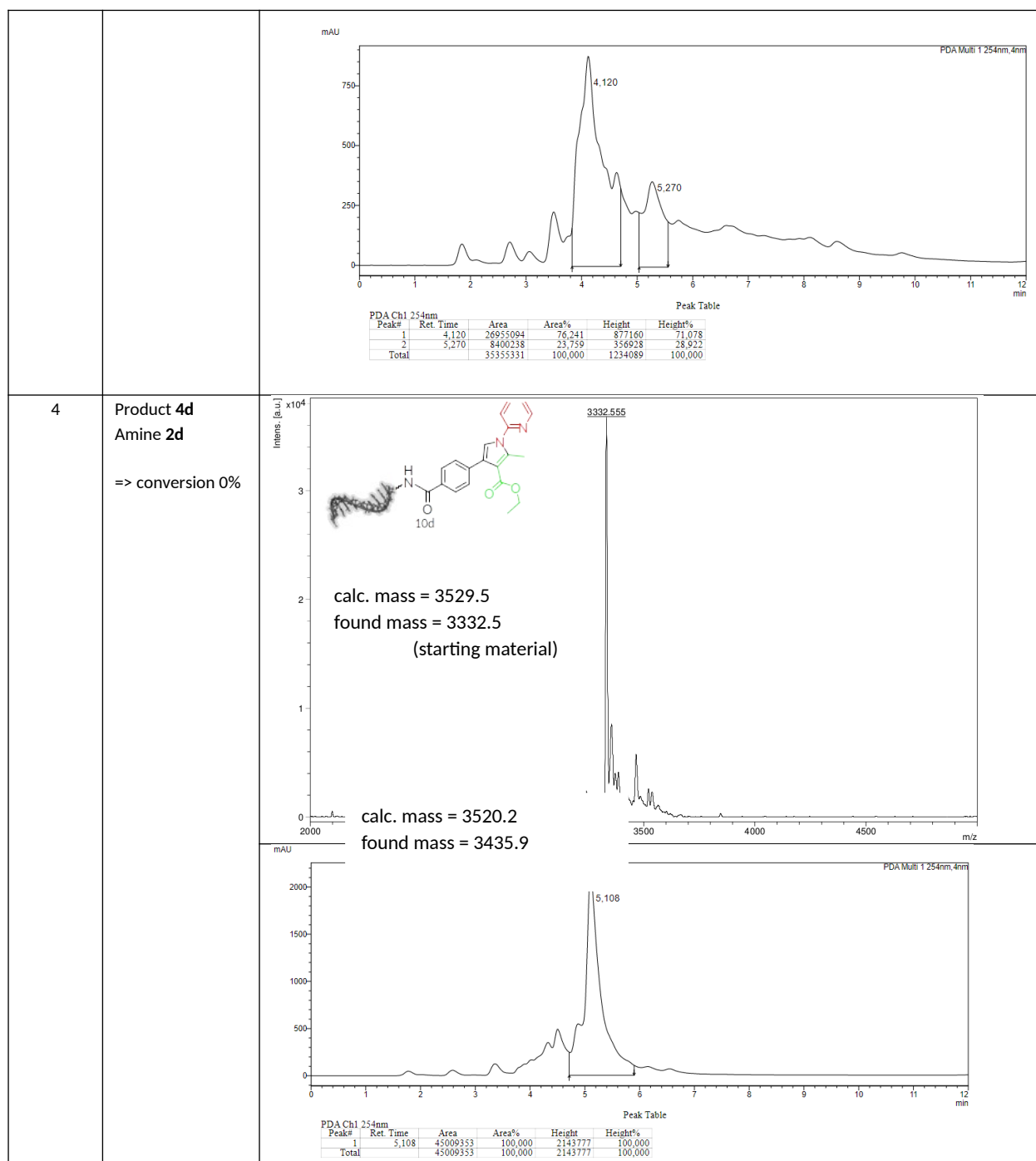
Table 23: Amine scope for the pyrrole synthesis on the CPG- bound DNA-aldehyde conjugate **1** according to GP2.^a



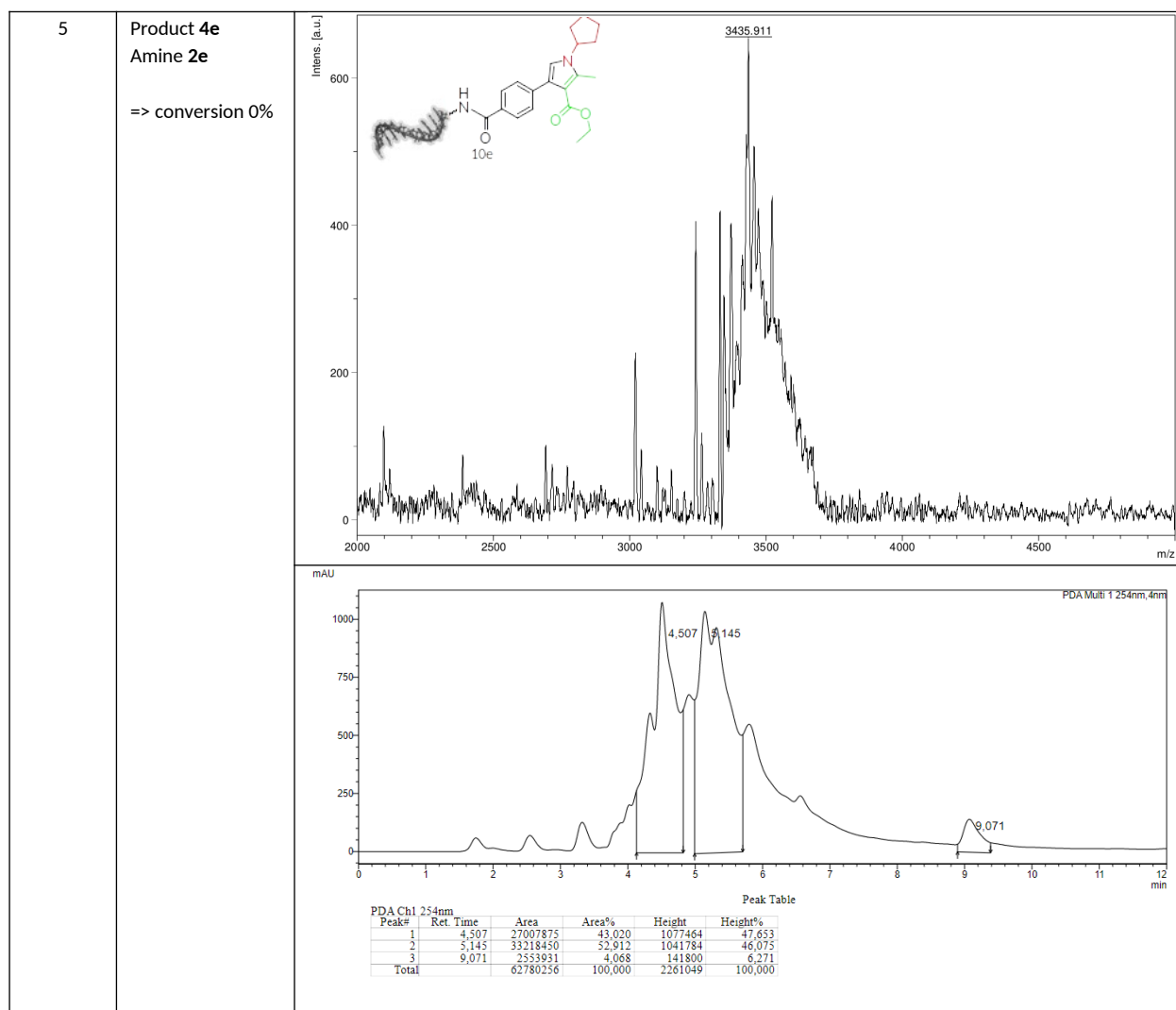
Experimental part



Experimental part



Experimental part

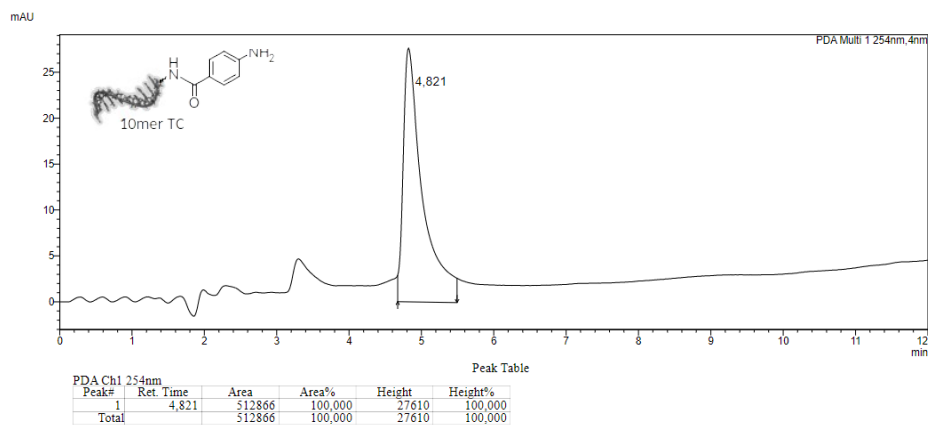


^a The suspension of DNA-aldehyde conjugate **1** (20 nmol), amine **2** (250 μ M), ethylcetoacetate **3** (250 μ M) and FeCl_3 (25 μ M) in MeNO_2 (40 μ L) was shaken at 80 $^\circ\text{C}$ for 6 h. DNA cleavage with 32% aq. Ammonia solution at ambient temperature for 4 h. ^b Determined by RP-HPLC analysis based on the ratios of **1** to **4**. ^c Determined by RP-HPLC analysis in ratio to the purity of the CPG-coupled DNA-aldehyde conjugate **1**. MeNO_2 = nitromethane. 10mer ATGC = 5'-NH₂-C6-GTCATGATCT-3'-CPG.

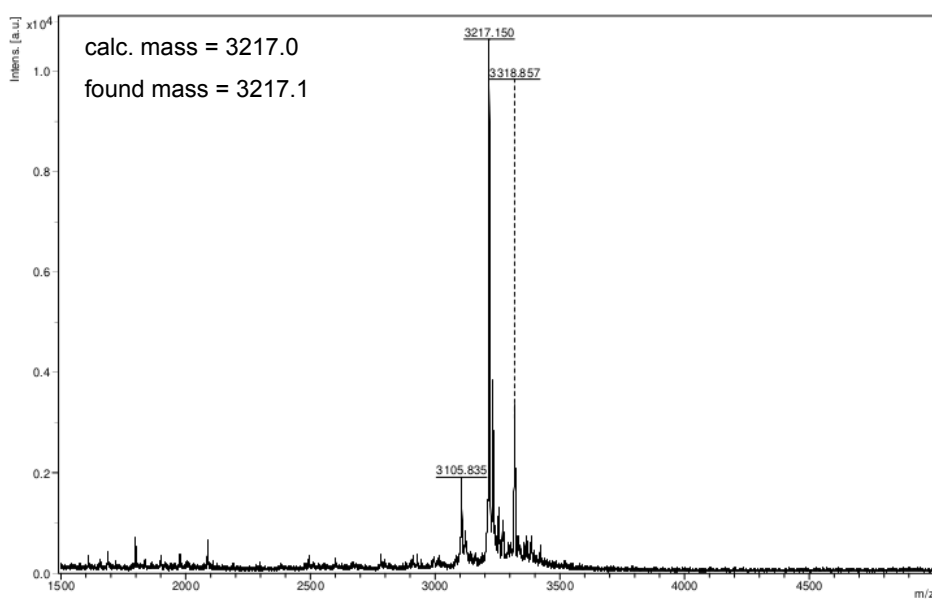
Experimental part

DNA conjugate 5: CPG-bound 10mer TC-(CH₂)₆-NH₂ conjugate was reacted with 4-amino-benzoic acid according to **GP1**.

HPLC trace of crude reaction mixture **5** (Analytical RP-HPLC).



MALDI-MS spectrum of crude reaction mixture **5**



Experimental part

Table 24: Test reaction for pyrrolidine synthesis on a CPG-bound pyrimidine DNA-amine conjugate **5** according to GP3.^a

Entry	Reaction conditions	MALDI spectrum of the crude reaction mixture
1	thioester 6 (200 mM) benzaldehyde 7 (200 mM) <i>Et₂NH</i> (200 mM) => conversion n.d. ^b	<p> calc. mass = 3484.0 found mass = 3271.9 (side product 9) </p> <p>10mer TC</p>
2	thioester 6 (200 mM) benzaldehyde 7 (200 mM) <i>MgI₂</i> (200 mM) => conversion n.d. ^b	<p> calc. mass = 3484.0 found mass = 2982.3 (side product 10) </p> <p>10mer TC</p>

^a The suspension of DNA-amine conjugate **5** (20 nmol), thioester **6** (200 mM) and benzaldehyde **7** (200 mM) and metal salt (200 mM) in DCE (40 μ L) was shaken at ambient temperature for 1 h. DNA cleavage with 32% aq. ammonia solution at ambient temperature for 30 min. ^b Determined by MALDI-TOF analysis, DCE = dichloroethane. n.d. = not detected, 10mer TC = 5'-TTC CTC TCC T-3'-CPG.

9.2 Building blocks selection

As the KNIME workflow for building block selection was adapted to the molecules class according to the considered library or to different requirements for designing each library, the most utilized nodes are reported.

- *SDF Reader*, to read the catalogue from which the building blocks were selected, such as the Enamine REAL space [166] or the Aldrich Market Select [37].
- *RDKit Functional Group Filter*, to divide the data set according to functional group. In this case, mono functional or bifunctional building blocks could be selected depending on which cycle of the library was being designed.
- *RDKit Molecule Catalogue Filter*, to filter out PAINS (Pan-Assay INterference compounds). In this node the three groups PAINS-A, -B and -C were selected.
- *MarvinSketch / RDKit Molecule Substructure Filter*, to filter specific unwanted substructures. Building blocks containing sulfur or bulky substituents in *ortho*-position to the reacting functional group were excluded with this method. The same combination of nodes was used to constrain the rings size or the presence of heteroatoms in rings.
- *RDKit R-Group Decomposition* as an alternative to the previous combination of nodes, to exclude fragments with hindered functional groups. By splitting the molecules in MCS (Multiple Common Substructure - defined by the *MarvinSketch* node) and R-groups, this method allowed for identification of each substituent in each position. After that, building blocks with bulky substituents in *ortho*-position could be eliminated. Additionally, the set of molecules could be refined according to activating or deactivating groups affecting the reactivity in the specific reaction. For example, if the initial functional moiety was a deactivating group, we ensured that the *meta*-position of all fragments was free.
- *RDKit Descriptor Calculation / Row Splitter*, to calculate specific molecular features such as molecular weight and trim accordingly. The molecular weight threshold was kept to 200 Da per each class of fragments per each cycle not to grow the molecules too much beyond the Lipinski rule of five. For aromatic structures, the number of aromatic rings was maintained to only one in order to not generate very planar and rigid molecules.
- *RDKit Diversity Picker*, to narrow down the size of the selection for big data set of building blocks without losing diversity.

For the library enumeration the following nodes were employed:

- *MarvinSketch*, to draw the reaction scheme. In this passage, the scheme was kept as minimal as possible and special attention was put into mapping reacting atoms in reactants and products. The output format was set to RXN.
- *RDKit Two Component Reaction*, to perform the reactions according to the previously sketched scheme. For the combinatorial nature of DELs, in this node the options “Uniquify products” and “Do matrix expansion” were flagged. This node was used for each step of the library synthesis and, in the case of MCRs (Multi Component Reactions), the reaction was split in additional steps.
- *Constant Value Column / Counter Generation / Column Aggregator*, to label the building blocks before the reaction (Figure SX). This step was essential to track back the products after all the reaction steps. Alternatively, an initial *Structure sheet* could be used where each building block per each step was listed with the respective SMILES string, label and code.
- *MolConverter / RDKit From Molecule / Joiner*, to homogenize the formats before and after the reaction and append the label to the respective fragment. This sequence was repeated for each step in order to label the final product as well.
- *RDKit Descriptor Calculation*, to compute the molecular properties of the virtually synthesized libraries. With this node, the Lipinski descriptors could be calculated such as the MW, SLogP, TPSA (Topological Polar Surface Area), NumHBD (Number of Hydrogen Bond Donors) and NumHBA (Number of Hydrogen Bond Acceptors). Alternatively, the 42 MQN descriptors could be calculated, which consider all topological and pharmacophoric molecular properties.
- *Constant Value Column*, to differently label the library members and the catalog used for comparison. In this way, the two data sets were colored differently in the scatter plot representing the chemical space and in the PMI triangle plot representing the shape diversity.
- *RDKit Generate Coords / Principal Moment of Inertia (PMI)-Derived Properties / PMI Triangle Scatter Plot*, to visualize trends in molecular shape within the libraries. The 3D coordinated were generated and they were used to calculate the PMI properties PMI 1, 2 and 3 and the npr 1 and 2.
- *Normalizer / PCA / 2D/3D Scatterplot*, to visualize the library member in the 3D chemical space. At first, the MQN descriptors values were normalized via the z-score method and then the PCA (Principal Component Analysis) was performed to reduce the dimensions of the

database to three, preserving more than 97% of information. Finally, the three PCA dimensions were used as axes in the scatter plot defining the chemical space. In the final scatter plot the color value was set according to the affiliation with the different data set, the library or the database for comparison.

This procedure was applied to the virtual generation of libraries based on Povarov, *aza*-Diels-Alder and Biginelli scaffolds. The three libraries were initialized with the same set of aldehyde-acid building blocks comprising 490 molecules and the rest of the substructures were selected according to the reaction. For the Povarov library 1000 anilines were diversely picked after filtering them by molecular weight and sterical hindrance at the *ortho*-position, while the dihydropyrrole was kept constant. For the *aza*-Diels-Alder library 1000 amines and anilines were diversely picked and the Danishefsky's diene was kept constant. For the Biginelli library 319 Ureas and 100 α -ketoesters were selected. The final sizes of the Povarov, *aza*-Diels-Alder and Biginelli libraries were approximately 340000, 310000 and 380000 respectively. Finally, they were compared with drug-like molecules in the Enamine REAL space, which at that time comprised 24 million compounds filtered by the Lipinski rule of five and Veber criteria [104]. From the database only 500,000 entries were randomly picked to scale it to the size of the three libraries. [22]

Additionally, prior to library synthesis, building blocks were selected for the first cycle of a library based on reductive amination and Suzuki coupling. This specific library was initiated by coupling a diamine, containing a *t*-*boc* protecting group on one of the two secondary amines, to the DNA tag. In this case, the selection involved fragments from the Enamine REAL space catalogue containing an aldehyde moiety for the reductive amination with the free secondary amine and a Bromine or Iodine for the Suzuki coupling with the boronic acid in the second cycle. With the combination of the above mentioned nodes a set of five aldehyde-halide fragments were selected and were reacted with 17 boronic acids. The last step was the substitution of the *t*-*boc* group with 100 carboxylic acids to introduce the third diversity point. This process produced an 8500 membered library which was, then, compared with the Enamine diverse REAL catalogue of drug-like compounds in terms of shape and chemical space coverage. This dataset contains roughly 38 million diverse compounds which comply with the rule of five and the Veber's criteria [101], [104] and does not contain PAINS or toxic compounds. For this purpose of comparing it to the library, the catalogue was trimmed to ca. 104,000 compounds filtering by molecular weight in the 400-550 range and then 10,000 entries were diversely picked with the appropriate node. Additionally, three Enamine databases were utilized:

Experimental part

- the BioReference Library, containing 2100 compounds with a broad biological activity, including almost 700 FDA approved drugs.
- the FDA approved drugs collection, used for completeness, containing 7 more FDA approved molecules.
- lead-like molecules with the following characteristics: $MW \leq 460$, $-4 \leq \text{SlogP} \leq 4.2$, $\text{HBA} \leq 9$, $\text{HBD} \leq 5$, $\text{rings} \leq 4$, $\text{rotatable bonds} \leq 10$.
- natural products - like compounds.

9.3 Hits validation by molecular docking

9.3.1 KNIME workflow

The enrichment factor (EFs) calculations were performed by Nils Yannik Schüssler, master student in Prof. Fried group at the Statistics department of the TU Dortmund, and the outcome consisted in an Excel file made of as many sheets as the number of proteins. In the first screening campaign 16 targets were screened, but for developing the KNIME workflow and for improving hit identification by docking only the three following proteins were considered: MKK7, MDM2, BCL-X_L.

Each sheet (one for each protein) of the excel file consisted of the following columns:

ProteinCode_ID, an identification number per protein

Region1Code_ID, an identification string per each cycle 1 building block

Region2Code_ID, an identification string per each cycle 2 scaffold

Region3Code_ID, an identification string per each cycle 3 carboxylic acid

reads.protein, number of all detected proteins

reads.control, number of control experiments

freq.protein, frequency of the considered protein

freq.control, frequency of the respective control

promille.protein, parts per thousand connectivity of the protein of interest

promille.control, parts per thousand connectivity of the respective control

enrichmentFactor, frequency of each sequence over the rank abundance

rank.Abundance, count of each unique sequence

Control, experiments used as control with which the enrichment factor was calculated

The workflow to read the above mentioned excel file is depicted in Figure 58 and started with an *Empty Table Creator* node followed by a *Counter Generation* and a *Table Row To Variable Loop Start* node. This combination of nodes allowed for reading iteratively all sheets, and consequently all proteins, of the excel file. Then, the variable output port of the *Table Row to Variable Loop Start* node was connected to the *Excel Reader (XLS)* node, whose input was set to be the excel file and in the flow variable section the SHEET_NAME parameter was controlled by the variable *Counter*.

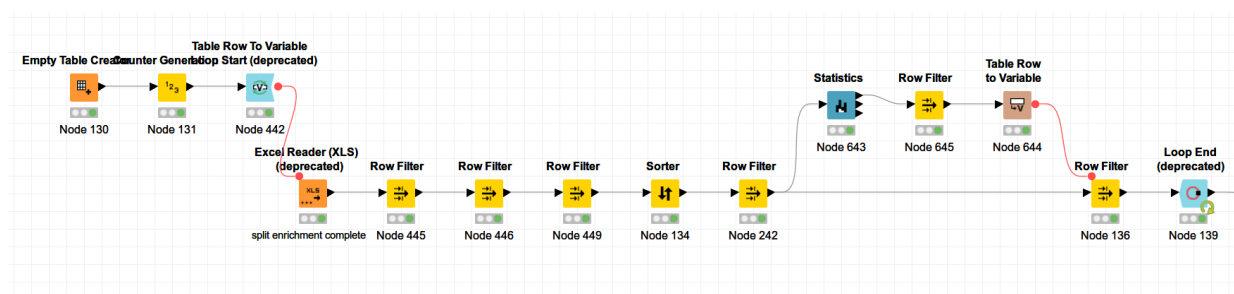


Figure 58: Input workflow for a preliminary filtering of the sequencing data.

The *Excel Reader* was followed by three *Row Filter* nodes which excluded the control experiments calculations from the analysis. Subsequently, a *Sorter* node ordered the *enrichmentFactor* column in a descending way. An additional *Row Filter* excluded the control *Streptavidine Beads* from the *Control* column and was connected to the *Statistics* node. This node calculated statistical parameters per each column of the input table transposing the columns into rows. The first output of this node was filtered to include only the *enrichmentFactor* row and this row was transformed into a variable via the *Table Row to Variable* node. The mean value used in the following *Row Filter* node as variable for controlling the lower bound parameter. The whole process was iterated for the rest of the proteins via a *Loop End*. In this way, the sequences and by consequence the molecules showing an EF higher than the average EF per each proteins were selected.

The next section of the workflow was necessary for appending the molecules structures to the table. First of all the *Excel Reader* node was used to open the Structure sheet, a file where the SMILES, codes and names were present per each building block and protein. The respective SMILES strings for the three cycles molecules were converted to structures with the *Molecule Type Cast* nodes and the structures were appended via three *Joiner* nodes. Additionally, the protein name was extracted from the Structure Sheet and joined to the table as well (Figure 59).

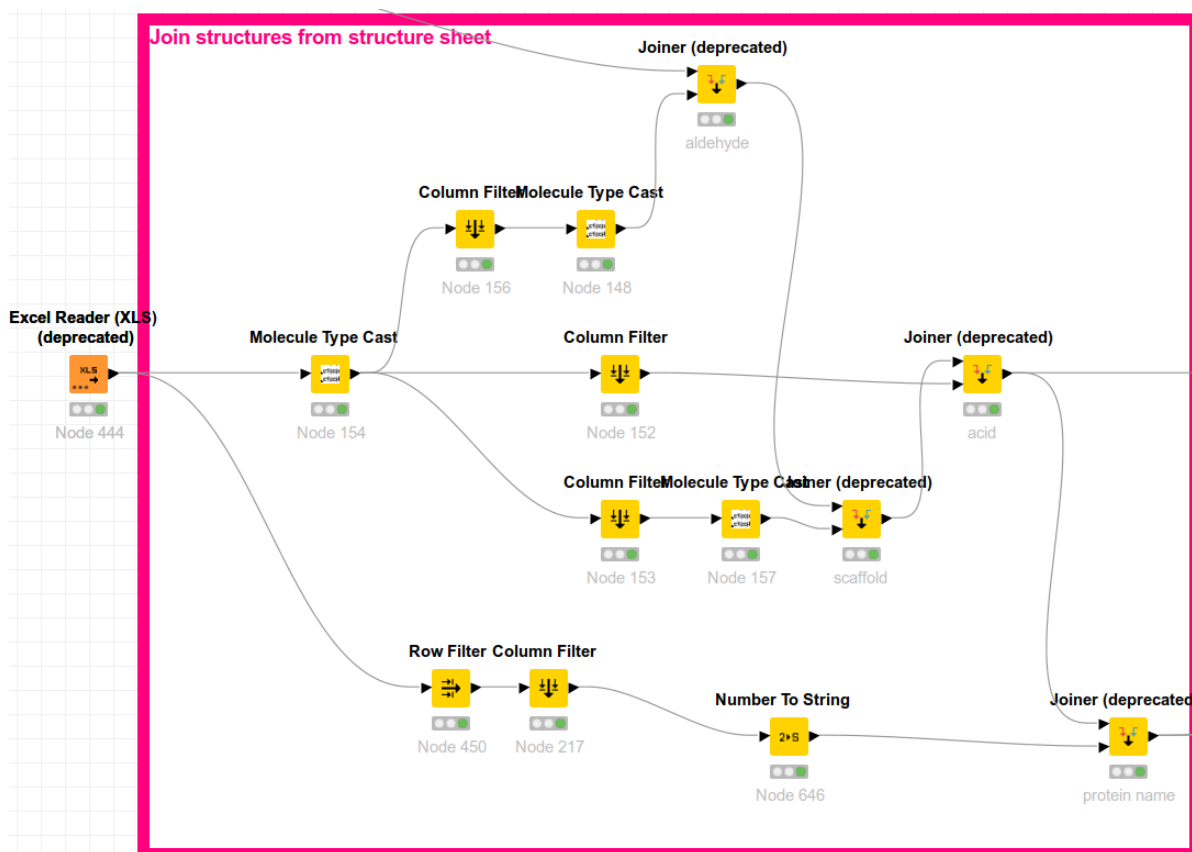


Figure 59: Workflow to assign the chemical structures to the respective labels.

Hence, the so-called frequent hitters were excluded from the data, as they are unspecific (Figure 60). For this purpose, a *GroupBy* node grouped the three cycles IDs and counted the unique proteins IDs. After that, the *Row Splitter* node excluded the combinations that bound more than four proteins. The second output of the *Row Splitter* was connected to the second input ports of three *Reference Row Splitter* nodes, one for each cycle. The first input port of the first node was fed by the *Joiner* that appended the protein name, and the columns to match were set to the *Region1Code_ID*, namely the cycle 1 aldehyde. The first output port of this node was connected to the second *Reference Row Splitter* node and this time the scaffold ID was matched. The same procedure was repeated for the cycle 3 carboxylic acid. At the end, an additional *Reference Row Splitter* node matched the RowIDs to select the entries from the joined table containing the at the same time frequent cycle 1, 2 and 3 molecules.

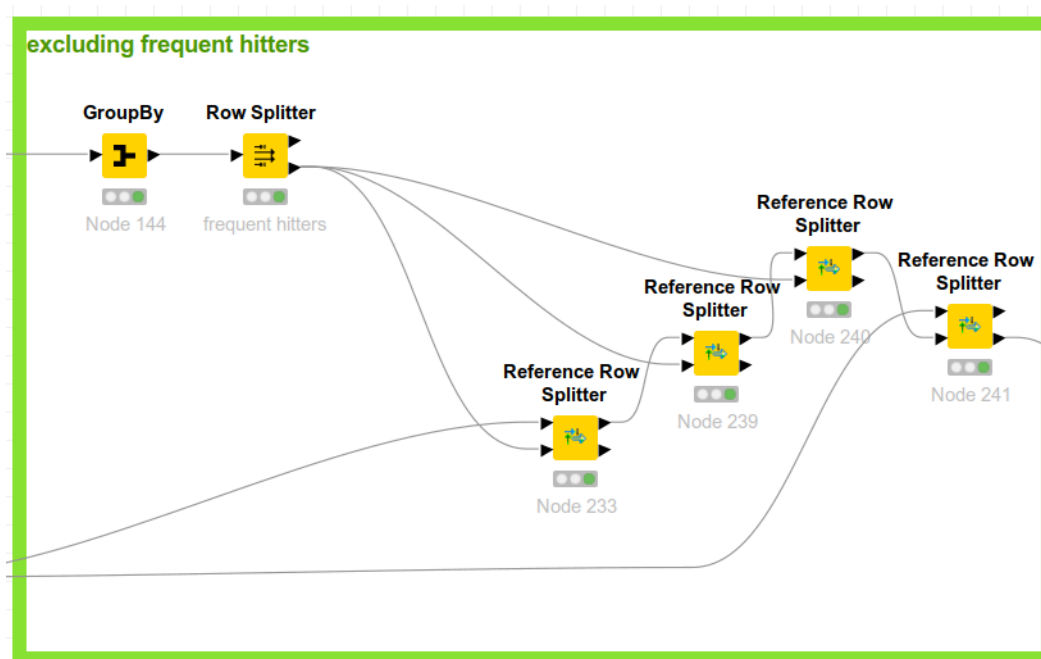


Figure 60: Workflow to exclude the frequent hitters: molecules which bound more than four targets.

At this point, the data were split according to the control experiments and the control *Amplified DEL* was considered, because more relevant to our analysis. To label the combinations (Figure 61) a *Constant Value Column* was used which appended the column "Comb_ID" with value "comb". Then, the *Counter Generation* node was employed to assign an integer to each row and a *Column Aggregator* to concatenate the constant value and the counter, in order to assign to each combination a unique ID. As some combination bound different proteins some duplicates with different IDs were present. This problem was addressed by employing a *Duplicate Row Filter* which was set to create an additional column with the ID of the duplicate. The duplicates were named after the respective reference combinations via a *Row Splitter*, a *Column Filter* and a *Column Rename* nodes, so that each combination showed a unique ID. After that, the columns created by the *Duplicate Row Filter* node were excluded by a *Column Filter* node.

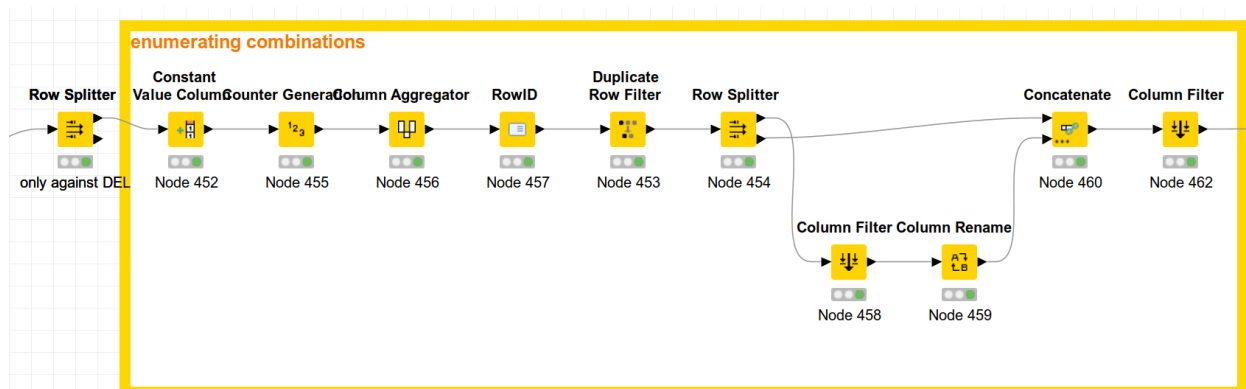


Figure 61: Workflow for labelling the combinations to track the molecules in the future steps.

The cycle 1, 2 and 3 molecules were then connected to form the complete structures. To do so, a loop through the targets was initialized by a *Group Loop Start*. Firstly, the cycle 1 aldehydes, presenting a carboxylic acid moiety for coupling to the DNA, were submitted to the *RDKit Chemical Transformation* node which generated an thioester on the carboxylic acid, to differentiate it from any other substructure in the dataset that could interfere with the product formation. For this purpose, the *MarvinSketch* node employed. Subsequently, the scaffold SMILES were modified via a *String Manipulation* with the expression: "replace(\$Scaffold_Smiles\$, "[R]", "*")" to be converted in structure with an exit vector via the *Molecule Type Cast*. The exit vector was then substituted by the cycle 1 aldehydes via a *RDKit Two Component Reaction* node and the right formed product was selected via visual inspection of the first rows of the table. As the correct product was differentiated by the RowID sequence number, the *RowID* node was used to translate the RowIDs into a column and the Row Filter node include only the rows containing the pattern "*_*_1_*". The output of the RowID node fed the second input port of the *RDKit Two Component Reaction* node, whose first output was connected with the initial table coming from the *Group Loop Start* node to retrieve the structures of the cycle 3 carboxylic acids. The last input port of the *RDKit Two Component Reaction* node received the amide bond formation reaction in RXN format from the *MarvinSketch* node. As in the case of the Povarov scaffold two amino-groups were available for amide bond formation, the *GroupBy* node was needed to filter the undesired product. The thioester linker was substituted by an amide bond with the *RDKit Chemical Transformation* node and the respective reaction drawn in *MarvinSketch*. The formed products were then to be labelled again with their proper combination IDs, so a *Column Filter* node from the initial *Group Loop Start* node included only the relevant columns and they were appended to the products by a *Column Appender* node. The last important step of this process was to save the structures in SDF format in order to export them to the docking software. To do so, the *Column*

Appender node was connected to the *Table Row to Variable* node and the column containing the protein name was converted into variable. Then, the path of the file was added via a *String Manipulation (Variable)* node and its output port was connected to the variable input port of a *SDF Writer* node, which wrote the SDF file in the proper folder with the proper name. The loop was then closed by a *Loop End* node in order to repeat the same process per each and every target protein (Figure 62).

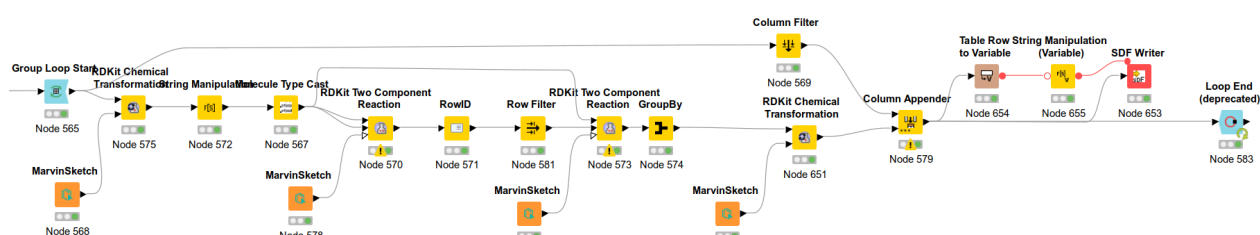


Figure 62: Workflow for the chemical reactions to form the final products.

9.3.2 Molecular Docking

9.3.2.1 Docking with SeeSAR

Initially, the proteins and the ligands were prepared in SeeSAR according to default settings and then the docking procedure was applied via the appropriate function. The molecules that were predicted to bind in a nmolar to molar range were exported and compared within KNIME.

9.3.2.2 Docking with Glide

Where not explicitly stated, the software settings were left unchanged. The protein structures were retrieved from the Protein Data Bank (PDB) using the codes shown in Table 25 via the *Get PDB* function of the software Maestro.

Table 25: Target proteins with respective PDB codes.

Target protein	PDB code
BCL-X _L	2YXJ
MDM2	MDN4
MKK7	6YG7

Subsequently, the protein crystallographic structures were prepared via the *Protein Preparation Wizard*. In particular, the missing loops were filled with *Prime*, while other species, including the co-crystallized ligands, were removed. This step allowed us to perform a blind docking, without biasing the

calculation with pre-existing constraints. The charges were calculated for $\text{pH}=7.2 \pm 0.2$. The remaining settings were left as default including the radius at which the water molecules were considered. Finally, the proteins were minimized according to the OPLS4 force field. The molecules generated by the KNIME workflow were prepared with the *LigPrep* function in Maestro. The charges were calculated for $\text{pH}=7.2 \pm 0.2$ and 32 possible conformations per ligand were generated. The prepared proteins were submitted to the *SiteMap* protocol, which searched for druggable pockets over the proteins surface area. Very small pockets were discarded due to the size of the molecules output of the library. Each pocket identified by the *SiteMap* was used as centre in the *Glide Grid Generation* protocol. In this case, the grid box was expanded to 20 Å in each direction. For the docking procedure, each site-based grid was employed sequentially and the *Glide Docking* function was performed at a standard precision level (SP). The number of poses per ligand was set to 50 and the strain energy for ligands was applied. The output of the docking protocol was exported in a table to be used in KNIME and being compared with the Enrichment Factors and the calculated affinity by SeeSAR. For this purpose, the *Pareto Ranking* node in KNIME was used, minimizing the docking score and the calculated affinity and maximizing the enrichment factor.

10 Abbreviations

ΔG	increment in Gibbs energy
A	adenosine
Ala	alanine
AMA	aqueous ammonia (30%) / aqueous methylamine (40%), 1:1, vol/vol
AMS	aldrich market select
Arg	arginine
B	Biginelli scaffold
BAK	BCL2 antagonist/killer
BAX	BCL2 associated X
BB	building block
BCL-X _L	B cell lymphoma X _L
BCL2	B cell lymphoma 2
BIOS	biology-oriented synthesis
boc/t-boc	tert-butyloxycarbonyl
C	cytidine
CLogP	calculated LogP
CPG	controlled pore glass
CPU	central processing unit
DA	aza-Diels-Alder scaffold
DCE	dichloroethane
DCM	dichloromethane
DEL	DNA-encoded library
DIPEA	N,N-Diisopropylethylamine
DMA	dimethylacetamide
DMF	dimethylformamide
DMSO	dimethyl sulphoxide
DNA	deoxyribonucleic acid
DOS	diversity-oriented synthesis

Abbreviations

EF	enrichment factor
EtOH	ethanol
FDA	food and drug administration
FF	force field
G	guanosine
Glu	glutamate
HATU	Hexafluorophosphate Azabenzotriazole Tetramethyl Uronium
HBA	hydrogen bond acceptor
HBD	hydrogen bond donor
HC	hierarchical clustering
HTS	high-throughput screening
JNK	c-Jun N-terminal kinase
KNIME	konstanz information miner
LG	leaving group
MALDI-TOF	Matrix-Assisted Laser Desorption/Ionization-Time Of Flight
MCR	multi-component reaction
MCS	multiple common substructure
MDM2	murine double minute 2
MeCN	acetonitrile
MeOH	methanol
MKK7	mitogen-activated protein kinase kinase 7
MOMP	mitochondria outer membrane permeabilisation
MQN	molecular quantum number
MW	molecular weight
NGS	next generation sequencing
nM	nanomolar
NP	natural product
OBOC	one-bead-one-compound
ORD	open-reaction database
P	Povarov scaffold

Abbreviations

PAINS	pan-assay interference compounds
PC	partitional clustering
PCA	principal component analysis
PCR	polymerase chain reaction
PDB	protein data bank
PEG	polyethylene glycole
PMI	principal moment of inertia
PPI	protein-protein-interaction
RA	rank abundance
RO5	rule of five
RotBonds	number of rotatable bonds
RP-HPLC	reverse phase – high performance liquid chromatography
rxns	reactions
Sc	silhouette coefficient
SF	scoring function
SLogP	hybrid LogP
SMARTS	SMILES arbitrary target specification
SMILES	simplified molecular input line system
T	thymidine
TEA	triethylamine
THF	tetrahydrofurane
THQ	tetrahydroquinoline
TPSA	topological polar surface area
VS	virtual screening
μM	micromolar

11 Acknowledgment

Among all the people I am grateful to for being part of my life, I focus here on the ones who played an important role in my doctoral studies. Firstly, I would like to thank Andreas Brunschweiger for assigning me this position in such interdisciplinary and stimulating research projects. He guided me through the doctoral studies, staying always patient and positive. Despite my research focus strayed from his background knowledge, he supported me with external collaborations, ensuring my growth as a scientist. I will never thank him enough for believing in me.

Although she is not part of the group anymore, I thank Mateja Klika Škopić for her help in every small task in the laboratory. I considered her to be the mother of the group, being always patient and caring. Even in her most stressful moments, she was always there to listen and give precious advises.

I thank Katharina for sharing ups and downs of these doctoral studies with me, for being always there with her smile and for her patience with the never ending translations from German to English.

A huge thank is deserved by Maria Sergani, our beloved secretary, who helped me so much, especially at the beginning of this PhD. Without her I would have been completely lost.

I thank Avinash, Elena and Suzanne for joining this group, making the atmosphere funnier and more relaxed. I especially thank Avinash and Elena for reading and correcting parts of this thesis.

I am thankful to this PhD programme because it allowed me to encounter two very important people in my life: Irene and Daniel. Irene showed me how enthusiastic a good scientist could possibly be. With her indestructible determination and discipline, she will always represent a model to me. Daniel, since the moment we've met for the first time, released the burden of hard moments during the course of these studies. Every obstacle which appeared to me as an insurmountable mountain, became a grain of sand to be blown away after talking with him.

I am thankful for Mais' friendship, which was born at a German course at the university and became so important to me to the point of defining her "my sister in Germany".

Finally, I am extremely grateful to my parents for their trust and their constant care. They never stopped believing in me, even when I had stopped myself. They are the special kind of parents to whom a daughter can tell anything: they would always listen and do their best to solve problems, even from another country! Words can not express my gratitude.

12 References

- [1] J. A. DiMasi, H. G. Grabowski and R. W. Hansen, *J. Health Econ.*, **2016**, 47, 20–33.
- [2] R. S. Bohacek, C. McMartin and W. C. Guida, *Med. Res. Rev.*, **1996**, 16, 3–50.
- [3] B. J. Bender, S. Gahbauer, A. Luttens, J. Lyu, C. M. Webb, R. M. Stein, E. A. Fink, T. E. Balius, J. Carlsson, J. J. Irwin and B. K. Shoichet, *Nat. Protoc.*, **2021**, 16, 4799–4832.
- [4] J.-L. Reymond, *Acc. Chem. Res.*, **2015**, 48, 722–730.
- [5] T. L. Foley, W. Burchett, Q. Chen, M. E. Flanagan, B. Kapinos, X. Li, J. I. Montgomery, A. S. Ratnayake, H. Zhu and M.-C. Peakman, *SLAS Discovery*, **2021**, 26, 263–280.
- [6] S. Brenner, R. A. Lerner, *Proc. Natl Acad. Sci. USA*, **1992**, 89, 5381–5383.
- [7] E. J. Corey, *Pure and Appl. Chem.*, **1967**, 14, 19–38.
- [8] A. L. Satz, A. Brunschweiler, M. E. Flanagan, A. Gloger, N. J. V. Hansen, L. Kuai, V. B. K. Kunig, X. Lu, D. Madsen, L. A. Marcaurette, C. Mulrooney, G. O'Donovan, S. Sakata and J. Scheuermann, *Nat. Rev. Methods Primers*, **2022**, 2, 3.
- [9] T. Hu, N. Chitnis, D. Monos and A. Dinh, *Hum. Immunol.*, **2021**, 82, 801–811.
- [10] J. C. Faver, K. Riehle, D. R. Lancia, J. B. J. Milbank, C. S. Kollmann, N. Simmons, Z. Yu and M. M. Matzuk, *ACS Comb. Sci.*, **2019**, 21, 75–82.
- [11] V. B. K. Kunig, M. Potowski, M. Akbarzadeh, M. Klika Škopić, D. Santos Smith, L. Arendt, I. Dormuth, H. Adihou, B. Andlovic, H. Karatas, S. Shaabani, T. Zarganes-Tzitzikas, C. G. Neochoritis, R. Zhang, M. Groves, S. M. Guéret, C. Ottmann, J. Rahnenführer, R. Fried, A. Dömling and A. Brunschweiler, *Angew. Chem.*, **2020**, 132, 20518–20522.
- [12] T. Kodadek, *Chem. Commun.*, **2011**, 47, 9757.
- [13] M. A. Clark, R. A. Acharya, C. C. Arico-Muendel, S. L. Belyanskaya, D. R. Benjamin, N. R. Carlson, P. A. Centrella, C. H. Chiu, S. P. Creaser, J. W. Cuzzo, C. P. Davie, Y. Ding, G. J. Franklin, K. D. Franzen, M. L. Geffter, S. P. Hale, N. J. V. Hansen, D. I. Israel, J. Jiang, M. J. Kavarana, M. S. Kelley, C. S. Kollmann, F. Li, K. Lind, S. Mataruse, P. F. Medeiros, J. A. Messer, P. Myers, H. O'Keefe, M. C. Oliff, C. E. Rise,

References

- A. L. Satz, S. R. Skinner, J. L. Svendsen, L. Tang, K. van Vloten, R. W. Wagner, G. Yao, B. Zhao and B. A. Morgan, *Nat. Chem. Biol.*, **2009**, 5, 647–654.
- [14] M. Wichert, N. Krall, W. Decurtins, R. M. Franzini, F. Pretto, P. Schneider, D. Neri and J. Scheuermann, *Nat. Chem.*, **2015**, 7, 241–249.
- [15] P. Dickson and T. Kodadek, *Org. Biomol. Chem.*, **2019**, 17, 4676–4688.
- [16] M. Song and G. T. Hwang, *J. Med. Chem.*, **2020**, 63, 6578–6599.
- [17] K. Götte, S. Chines and A. Brunschweiler, *Tetrahedron Lett.*, **2020**, 61, 151889.
- [18] M. L. Malone and B. M. Paegel, *ACS Comb. Sci.*, **2016**, 18, 182–187.
- [19] P. R. Fitzgerald and B. M. Paegel, *Chem. Rev.*, **2021**, 121, 7155–7177.
- [20] R. D. Taylor, M. MacCoss and A. D. G. Lawson, *J. Med. Chem.*, **2014**, 57, 5845–5859.
- [21] J. D. Watson and F. H. C. Crick, *Nature*, **1953**, 171, 737–738.
- [22] M. Potowski, F. Losch, E. Wünnemann, J. K. Dahmen, S. Chines and A. Brunschweiler, *Chem. Sci.*, **2019**, 10, 10481–10492.
- [23] O. D’Augustin, S. Huet, A. Campalans and J. P. Radicella, *IJMS*, **2020**, 21, 8360.
- [24] J. Nielsen, S. Brenner and K. D. Janda, *J. Am. Chem. Soc.*, **1993**, 115, 9812–9813.
- [25] Y. Huang, Y. Li and X. Li, *Nat. Chem.*, **2022**, 14, 129–140.
- [26] B. Cai, D. Kim, S. Akhand, Y. Sun, R. J. Cassell, A. Alpsoy, E. C. Dykhuizen, R. M. Van Rijn, M. K. Wendt and C. J. Krusemark, *J. Am. Chem. Soc.*, **2019**, 141, 17057–17061.
- [27] L. K. Petersen, A. B. Christensen, J. Andersen, C. G. Folkesson, O. Kristensen, C. Andersen, A. Alzu, F. A. Sløk, P. Blakskjær, D. Madsen, C. Azevedo, I. Micco and N. J. V. Hansen, *J. Am. Chem. Soc.*, **2021**, 143, 2751–2756.
- [28] H. Deng, H. O’Keefe, C. P. Davie, K. E. Lind, R. A. Acharya, G. J. Franklin, J. Larkin, R. Matico, M. Neeb, M. M. Thompson, T. Lohr, J. W. Gross, P. A. Centrella, G. K. O’Donovan, K. L. (Sargent) Bedard, K. van Vloten, S. Mataruse, S. R. Skinner, S. L. Belyanskaya, T. Y. Carpenter, T. W. Shearer, M. A. Clark, J. W. Cuzzo, C. C. Arico-Muendel and B. A. Morgan, *J. Med. Chem.*, **2012**, 55, 7061–7079.
- [29] L. Kuai, T. O’Keefe and C. Arico-Muendel, *SLAS Discovery*, **2018**, 23, 405–416.
- [30] H. C. S. Chan, H. Shan, T. Dahoun, H. Vogel and S. Yuan, *Trends Pharmacol. Sci.*, **2019**, 40, 592–604.

References

- [31] F. K. Brown, *Annu. Rep. Med. Chem.*, **1998**, 33, 375–384.
- [32] J. Gasteiger and K. Funatsu, *J. Comput. Chem. Jpn.*, **2006**, 5, 53–58.
- [33] E. López-López, J. Bajorath and J. L. Medina-Franco, *J. Chem. Inf. Model.*, **2021**, 61, 26–35.
- [34] K. Martinez-Mayorga, A. Madariaga-Mazon, J. L. Medina-Franco and G. Maggiora, *Expert Opin. Drug Discov.*, **2020**, 15, 293–306.
- [35] D. Mendez, A. Gaulton, A. P. Bento, J. Chambers, M. De Veij, E. Félix, M. P. Magariños, J. F. Mosquera, P. Mutowo, M. Nowotka, M. Gordillo-Marañón, F. Hunter, L. Junco, G. Mugumbate, M. Rodriguez-Lopez, F. Atkinson, N. Bosc, C. J. Radoux, A. Segura-Cabrera, A. Hersey and A. R. Leach, *Nucleic Acids Research*, **2019**, 47, D930–D940.
- [36] <https://enamine.net/compound-collections/screening-collection> (accessed on 28.05.2022)
- [37] <https://www.sigmaaldrich.com/DE/en/services/custom-products/small-molecule-library-design> (accessed on 28.05.2022)
- [38] L. Ruddigkeit, R. van Deursen, L. C. Blum and J.-L. Reymond, *J. Chem. Inf. Model.*, **2012**, 52, 2864–2875.
- [39] T. Sterling and J. J. Irwin, *J. Chem. Inf. Model.*, **2015**, 55, 2324–2337.
- [40] D. Weininger, *J. Chem. Inf. Model.*, **1988**, 28, 31–36.
- [41] D. Weininger, A. Weininger and J. L. Weininger, *J. Chem. Inf. Comput. Sci.*, **1989**, 29, 97–101.
- [42] <https://www.daylight.com/dayhtml/doc/theory/theory.smarts.html> (accessed on 27.05.2022).
- [43] A. Saxena, M. Prasad, A. Gupta, N. Bharill, O. P. Patel, A. Tiwari, M. J. Er, W. Ding and C.-T. Lin, *Neurocomputing*, **2017**, 267, 664–681.
- [44] C. Fraley, *Comput. J.*, **1998**, 41, 578–588.
- [45] F. Murtagh, *Comput. J.*, **1983**, 26, 354–359.
- [46] S. Theodoridis, C. Rama, Academic Press, **2013**.
- [47] S. Ghosh, S. K. Dubey, *Int. J. Adv. Comput. Sci. Appl.*, **2013**, 4, 35–39.
- [48] J. Peng, W. Wang, Y. Yu, H. Gu and X. Huang, *CJCP*, **2018**, 31, 404–420.

- [49] M. R. Berthold, N. Cebron, F. Dill, T. R. Gabriel, T. Kötter, T. Meinl, P. Ohl, K. Thiel and B. Wiswedel, *SIGKDD Explor. Newsl.*, **2009**, 11, 26–31.
- [50] (A) RDKit: Open-source cheminformatics. <https://www.rdkit.org/>. (B) Indigo toolkit, GGA Software Services, <http://ggasoftware.com/>
- [51] C. Steinbeck, C. Hoppe, S. Kuhn, M. Floris, R. Guha and E. Willighagen, *CPD*, **2006**, 12, 2111–2120.
- [52] Marvin 16.7. 11, 2016, ChemAxon (<http://www.chemaxon.com>)
- [53] (A) <https://hub.knime.com/>; (B) <https://forum.knime.com/> (accessed on 29.05.2022)
- [54] A. J. Kooistra, M. Vass, R. McGuire, R. Leurs, I. J. P. de Esch, G. Vriend, S. Verhoeven and C. de Graaf, *ChemMedChem*, **2018**, 13, 614–626.
- [55] M. P. Mazanetz, R. J. Marmon, C. B. T. Reisser and I. Morao, *CTMC*, **2013**, 12, 1965–1979.
- [56] W. H. B. Sauer and M. K. Schwarz, *J. Chem. Inf. Comput. Sci.*, **2003**, 43, 987–1003.
- [57] V.B.K. Kunig, C. Ehrt, A. Dömling, A. Brunschweiler, *Org. Lett.*, **2019**, 21, 7238.
- [58] D. W. Piotrowski, P. Richardson, S. A. Green, R. A. Shenvi, J. S. Chen, P. S. Baran and P. E. Dawson, *J. Am. Chem. Soc.*, **2019**, 141, 9998–10006
- [59] M. Klika Škopić, K. Götte, C. Gramse, M. Dieter, S. Pospich, S. Raunser, R. Weberskirch and A. Brunschweiler, *J. Am. Chem. Soc.*, **2019**, 141, 10546–10555.
- [60] D. Chouikhi, M. Ciobanu, C. Zambaldo, V. Duplan, S. Barluenga and N. Winssinger, *Chem. Eur. J.*, **2012**, 18, 12698–12704.
- [61] A. Martín, C. A. Nicolaou and M. A. Toledo, *Commun. Chem.*, **2020**, 3, 127.
- [62] T. Klucznik, B. Mikulak-Klucznik, M. P. McCormack, H. Lima, S. Szymkuć, M. Bhowmick, K. Molga, Y. Zhou, L. Rickershauser, E. P. Gajewska, A. Toutchkine, P. Dittwald, M. P. Startek, G. J. Kirkovits, R. Roszak, A. Adamski, B. Sieredzińska, M. Mrksich, S. L. J. Trice and B. A. Grzybowski, *Chem.*, **2018**, 4, 522–532.
- [63] M. H. S. Segler, M. Preuss and M. P. Waller, *Nature*, **2018**, 555, 604–610.
- [64] Y. Gong, D. Xue, G. Chuai, J. Yu and Q. Liu, *Chem. Sci.*, **2021**, 12, 14459–14472.
- [65] V. Delannée and M. C. Nicklaus, *J. Cheminform.*, **2020**, 12, 72.
- [66] G. M. Ghiandoni, M. J. Bodkin, B. Chen, D. Hristozov, J. E. A. Wallace, J. Webster and V. J. Gillet, *J. Comput. Aided. Mol. Des.*, **2020**, 34, 783–803.

- [67] P. Schwaller, D. Probst, A. C. Vaucher, V. H. Nair, D. Kreutter, T. Laino and J.-L. Reymond, *Nat. Mach. Intell.*, **2021**, 3, 144–152.
- [68] <https://www.reaxys.com> (accessed on 24.05.2022).
- [69] <https://www.cas.org/cas-data/cas-reactions> (accessed on 25.05.2022).
- [70] D. M. Lowe, PhD diss., University of Cambridge, **2012**.
- [71] S. M. Kearnes, M. R. Maser, M. Wleklinski, A. Kast, A. G. Doyle, S. D. Dreher, J. M. Hawkins, K. F. Jensen and C. W. Coley, *J. Am. Chem. Soc.*, **2021**, 143, 18820–18826.
- [72] I. Mierswa, M. Wurst, R. Klinkenberg, M. Scholz and T. Euler, in Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining - KDD '06, ACM Press, Philadelphia, PA, USA, **2006**, 935.
- [73] M. R. Berthold, N. Cebron, F. Dill, T. R. Gabriel, T. Kötter, T. Meinl, P. Ohl, C. Sieb, K. Thiel and B. Wiswedel, *Springer*, **2007**.
- [74] N. G. Paciaroni, J. M. Ndungu and T. Kodadek, *Chem. Commun.*, **2020**, 56, 4656–4659.
- [75] W. Abdou, C. Bloch, D. Charlet and F. Spies, Evolutionary Computation in Combinatorial Optimization, eds. J.-K. Hao and M. Middendorf, Springer Berlin Heidelberg, **2012**, 7245, 194–205.
- [76] M. Klika Škopić, F. Losch, A. E. McMillan, N. Willeke, M. Malenica, L. Bering, J. Bode, A. Brunschweiler, *Org. Lett.* **2022**, accepted.
- [77] M. Feher and J. M. Schmidt, *J. Chem. Inf. Comput. Sci.*, **2003**, 43, 218–227.
- [78] N. Schneider, D. M. Lowe, R. A. Sayle, M. A. Tarselli and G. A. Landrum, *J. Med. Chem.* **2016**, 59, 4385–4402.
- [79] J. C. Bezdek, *J. Cybernetics*, **1973**, 3, 58–73.
- [80] M. Comiter, M. Cha, H. T. Kung and S. Teerapittayanon, in 2016 23rd International Conference on Pattern Recognition (ICPR), IEEE, Cancun, **2016**, 2331–2337.
- [81] S. Jeganathan, M. Tsukamoto and M. Schlosser, *Synthesis*, **1990**, 1990, 109–111.
- [82] J.-B. Feng and X.-F. Wu, *J. Heterocyclic Chem.*, **2017**, 54, 794–798.
- [83] C. Che, Z. Qian, M. Wu, Y. Zhao and G. Zhu, *J. Org. Chem.*, **2018**, 83, 5665–5673.
- [84] Z. Cao, H. Zhu, X. Meng, L. Tian, G. Chen, X. Sun and J. You, *J. Org. Chem.*, **2016**, 81, 12401–12407.

References

- [85] M. Hadjebi, M. S. Hashtroudi, H. R. Bijanzadeh and S. Balalaie, *HCA*, **2011**, 94, 382–388.
- [86] Y. Wan, R. Yuan, F.-R. Zhang, L.-L. Pang, R. Ma, C.-H. Yue, W. Lin, W. Yin, R.-C. Bo and H. Wu, *Synth. Commun.*, **2011**, 41, 2997–3015.
- [87] D.-Q. Shi, Y. Zou, Y. Hu and H. Wu, *J. Heterocyclic Chem.*, **2011**, 48, 896–900.
- [88] L.-Y. Zeng, B. Xi, K. Huang, J. Bi, L. Wei, C. Cai and S. Liu, *ACS Comb. Sci.*, **2019**, 21, 656–665.
- [89] H. Singh, J. Sindhu, J. M. Khurana, C. Sharma and K. R. Aneja, *Aust. J. Chem.*, **2013**, 66, 1088.
- [90] S. Steenken, S.V. Jovanovic, *J. Am. Chem. Soc.*, 1997, 119(3), 617–618.
- [91] N. Wongsu, U. Sommart, T. Ritthiwigrom, A. Yazici, S. Kanokmedhakul, K. Kanokmedhakul, A. C. Willis and S. G. Pyne, *J. Org. Chem.*, **2013**, 78, 1138–1148.
- [92] A. T. Khan, M. Lal, P. Ray Bagdi, R. Sidick Basha, P. Saravanan and S. Patra, *Tetrahedron Lett.*, 2012, 53, 4145–4150. (B) S. Maiti, S. Biswas and U. Jana, *J. Org. Chem.*, **2010**, 75, 1674–1683
- [93] W. Huang, J. Chin, L. Karpinski, G. Gustafson, C. M. Baldino, L.Yu, *Tetraedron Lett.*, **2006**, 47(28), 4911–4915.
- [94] M. Berthet, F. Davanier, G. Dujardin, J. Martinez and I. Parrot, *Chem. Eur. J.*, **2015**, 21, 11014–11016.
- [95] O. Eidam and A. L. Satz, *Med. Chem. Commun.*, **2016**, 7, 1323–1331.
- [96] J.-L. Reymond, R. van Deursen, L. C. Blum and L. Ruddigkeit, *Med. Chem. Commun.*, **2010**, 1, 30.
- [97] W. Warr, Report on an NIH Workshop on Ultralarge Chemistry Databases, *Chemistry*, **2021**.
- [98] R. J. Spandl, A. Bender and D. R. Spring, *Org. Biomol. Chem.*, **2008**, 6, 1149
- [99] W. Wilk, T. J. Zimmermann, M. Kaiser and H. Waldmann, *Biol. Chem.*, **2010**, 391, 491–497.
- [100] I. Pavlinov, E. M. Gerlach and L. N. Aldrich, *Org. Biomol. Chem.*, **2019**, 17, 1608–1623.
- [101] C. A. Lipinski, *Drug Discov.*, **2004**, 1, 337–341.
- [102] M.-Q. Zhang and B. Wilkinson, *Curr. Opin. Biotechnol.*, **2007**, 18, 478–488.

References

- [103] M. P. Pollastri, *Curr. Prot. Pharmacol.*, **2010**, 9.12.1-8.
- [104] P. Ertl, B. Rohde and P. Selzer, *J. Med. Chem.*, **2000**, 43, 3714–3717.
- [105] <https://enamine.net/compound-libraries/bioactive-libraries/fda-approved-drugs-collection> (accessed on 06.05.2022).
- [106] K. T. Nguyen, L. C. Blum, R. van Deursen and J.-L. Reymond, *ChemMedChem*, **2009**, 4, 1803–1805.
- [107] E. J. Martin, J. M. Blaney, M. A. Siani, D. C. Spellmeyer, A. K. Wong and W. H. Moos, *J. Med. Chem.*, **1995**, 38, 1431–1436.
- [108] J. J. Irwin, K. G. Tang, J. Young, C. Dandarchuluun, B. R. Wong, M. Khurelbaatar, Y. S. Moroz, J. Mayfield and R. A. Sayle, *J. Chem. Inf. Model.*, **2020**, 60, 6065–6073.
- [109] T. Hoffmann and M. Gastreich, *Drug Discov.*, **2019**, 24, 1148–1156.
- [110] A. Martín, C. A. Nicolaou and M. A. Toledo, *Commun. Chem.*, **2020**, 3, 127.
- [111] R. Pikalyova, Y. Zabolotna, D. M. Volochnyuk, D. Horvath, G. Marcou and A. Varnek, *Mol. Inform.*, **2022**, 41, 2100289.
- [112] Y. Zabolotna, D. M. Volochnyuk, S. V. Ryabukhin, K. Gavrylenko, D. Horvath, O. Klimchuk, O. Oksiuta, G. Marcou and A. Varnek, *J. Chem. Inf. Model.*, **2022**, 62, 2151–2163.
- [113] E. Lenci, L. Baldini and A. Trabocchi, *Bioorg. Med. Chem.*, **2021**, 41, 116218.
- [114] J. Baell and M. A. Walters, *Nature*, **2014**, 513, 481–483
- [115] F. W. Goldberg, J. G. Kettle, T. Kogej, M. W. D. Perry and N. P. Tomkinson, *Drug Discov.*, **2015**, 20, 11–17.
- [116] D. E. Patterson, R. D. Cramer, A. M. Ferguson, R. D. Clark and L. E. Weinberger, *J. Med. Chem.*, **1996**, 39, 3049–3059.
- [117] H. Abdi and L. J. Williams, *WIREs Comp. Stat.*, **2010**, 2, 433–459.
- [118] G. Schaftenaar and J. de Vlieg, *J. Comput. Aided Mol. Des.*, **2012**, 26, 311–318.
- [119] R. Mannhold, G. I. Poda, C. Ostermann and I. V. Tetko, *J. Pharmac. Sci.*, **2009**, 98, 861–893.
- [120] G. M. Keserü and G. M. Makara, *Nat. Rev. Drug Discov.*, **2009**, 8, 203–212.
- [121] X.-Y. Meng, H.-X. Zhang, M. Mezei and M. Cui, *CAD*, **2011**, 7, 146–157.
- [122] J. Li, A. Fu and L. Zhang, *Interdiscip. Sci. Comput. Life Sci.*, **2019**, 11, 320–328.

References

- [123] J.J. Irwin, D. M. Lorber, S. L. MCGovern, B. Wei, B. K. Shoichet, *Comput Nanosci Nanotechnol*, **2002**, 2, 50–51
- [124] W. L. Jorgensen, D. S. Maxwell and J. Tirado-Rives, *J. Am. Chem. Soc.*, **1996**, 118, 11225–11236.
- [125] J. Michel, J. Tirado-Rives and W. L. Jorgensen, *J. Phys. Chem. B*, **2009**, 113, 13337–13346.
- [126] S.-Y. Huang and X. Zou, *J. Chem. Inf. Model.*, **2010**, 50, 262–273.
- [127] E. Fischer, *Dtsch Chem. Ges*, **1894**, 27, 2984–2993
- [128] D. E. Koshland Jr, *Angew. Chem. Int. Ed.*, **2010**, 33, 2375–2378
- [129] SeeSAR version 12.0.1; BioSolveIT GmbH, Sankt Augustin, Germany, 2022, www.biosolveit.de/SeeSAR
- [130] Friesner, R. A.; Murphy, R. B.; Repasky, M. P.; Frye, L. L.; Greenwood, J. R.; Halgren, T. A.; Sanschagrin, P. C.; Mainz, D. T., *J. Med. Chem.*, **2006**, 49, 6177–6196
- [131] N. Schneider, G. Lange, S. Hindle, R. Klein and M. Rarey, *J. Comput. Aided Mol. Des.*, **2013**, 27, 15–29.
- [132] R. A. Friesner, J. L. Banks, R. B. Murphy, T. A. Halgren, J. J. Klicic, D. T. Mainz, M. P. Repasky, E. H. Knoll, M. Shelley, J. K. Perry, D. E. Shaw, P. Francis and P. S. Shenkin, *J. Med. Chem.*, **2004**, 47, 1739–1749.
- [133] K. Lansu, J. Karpiak, J. Liu, X.-P. Huang, J. D. McCorvy, W. K. Kroeze, T. Che, H. Nagase, F. I. Carroll, J. Jin, B. K. Shoichet and B. L. Roth, *Nat. Chem. Biol.*, **2017**, 13, 529–536.
- [134] J. P. Roney and S. Ovchinnikov, *Biophysics*, **2022**.
- [135] K. M. Ruff and R. V. Pappu, *J. Mol. Biol.*, **2021**, 433, 167208.
- [136] T. A. Halgren, *J. Chem. Inf. Model.*, **2009**, 49, 377–389.
- [137] P. Kolb and J. Irwin, *CTMC*, **2009**, 9, 755–770.
- [138] M. Lemke, H. Ravenscroft, N. J. Rueb, D. Kireev, D. Ferraris and R. M. Franzini, *Bioorg. Med. Chem. Lett.*, **2020**, 30, 127464.
- [139] Y. Huang, L. Meng, Q. Nie, Y. Zhou, L. Chen, S. Yang, Y.M.E Fung, X. Li, C. Huang, Y. Cao, Y. Li, *Nat. Chem.*, **2021**, 13, 77–88.
- [140] B. Dale, M. Cheng, K.-S. Park, H. Ü. Kaniskan, Y. Xiong and J. Jin, *Nat. Rev. Cancer*, **2021**, 21, 638–654.

References

- [141] A. Ashkenazi, W. J. Fairbrother, J. D. Levenson and A. J. Souers, *Nat. Rev. Drug Discov.*, **2017**, 16, 273–284.
- [142] S. Rajan, M. Choi, K. Baek and H. S. Yoon, *Proteins*, **2015**, 83, 1262–1272.
- [143] E. F. Lee, P. E. Czabotar, B. J. Smith, K. Deshayes, K. Zobel, P. M. Colman and W. D. Fairlie, *Cell Death Differ.*, **2007**, 14, 1711–1713.
- [144] T. Oltsdorf, S. W. Elmore, A. R. Shoemaker, R. C. Armstrong, D. J. Augeri, B. A. Belli, M. Bruncko, T. L. Deckwerth, J. Dinges, P. J. Hajduk, M. K. Joseph, S. Kitada, S. J. Korsmeyer, A. R. Kunzer, A. Letai, C. Li, M. J. Mitten, D. G. Nettesheim, S. Ng, P. M. Nimmer, J. M. O'Connor, A. Oleksijew, A. M. Petros, J. C. Reed, W. Shen, S. K. Tahir, C. B. Thompson, K. J. Tomaselli, B. Wang, M. D. Wendt, H. Zhang, S. W. Fesik and S. H. Rosenberg, *Nature*, **2005**, 435, 677–681.
- [145] L. Wang, G. A. Doherty, A. S. Judd, Z.-F. Tao, T. M. Hansen, R. R. Frey, X. Song, M. Bruncko, A. R. Kunzer, X. Wang, M. D. Wendt, J. A. Flygare, N. D. Catron, R. A. Judge, C. H. Park, S. Shekhar, D. C. Phillips, P. Nimmer, M. L. Smith, S. K. Tahir, Y. Xiao, J. Xue, H. Zhang, P. N. Le, M. J. Mitten, E. R. Boghaert, W. Gao, P. Kovar, E. F. Choo, D. Diaz, W. J. Fairbrother, S. W. Elmore, D. Sampath, J. D. Levenson and A. J. Souers, *ACS Med. Chem. Lett.*, **2020**, 11, 1829–1836.
- [146] S. Crespo-Garcia, P. R. Tsuruda, A. Dejda, R. D. Ryan, F. Fournier, S. Y. Chaney, F. Pilon, T. Dogan, G. Cagnone, P. Patel, M. Buscarlet, S. Dasgupta, G. Girouard, S. R. Rao, A. M. Wilson, R. O'Brien, R. Juneau, V. Guber, A. Dubrac, C. Beausejour, S. Armstrong, F. A. Mallette, C. B. Yohn, J.-S. Joyal, D. Marquess, P. J. Beltran and P. Sapieha, *Cell Metabol.*, **2021**, 33, 818–832.
- [147] S. Nag, J. Qin, K. S. Srivenugopal, M. Wang, R. Zhang, *J. Biomed. Res.*, **2013**, 27(4), 254–271.
- [148] B. Anil, C. Riedinger, J. A. Endicott and M. E. M. Noble, *Acta Crystallogr. D. Biol. Crystallogr.*, **2013**, 69, 1358–1366.
- [149] M. Konopleva, G. Martinelli, N. Daver, C. Papayannidis, A. Wei, B. Higgins, M. Ott, J. Mascarenhas and M. Andreeff, *Leukemia*, **2020**, 34, 2858–2874.
- [150] S. Baek, P. S. Kutchukian, G. L. Verdine, R. Huber, T. A. Holak, K. W. Lee and G. M. Popowicz, *J. Am. Chem. Soc.*, **2012**, 134, 103–106.
- [151] A. M. Bode and Z. Dong, *Mol. Carcinog.*, **2007**, 46, 591–598.

References

- [152] A. Shraga, E. Olshvang, N. Davidzohn, P. Khoshkenar, N. Germain, K. Shurrush, S. Carvalho, L. Avram, S. Albeck, T. Unger, B. Lefker, C. Subramanyam, R. L. Hudkins, A. Mitchell, Z. Shulman, T. Kinoshita and N. London, *Cell Chem. Biol.*, **2019**, *26*, 98-108.
- [153] R. Lonsdale and R. A. Ward, *Chem. Soc. Rev.*, **2018**, *47*, 3816–3830.
- [154] M. Schröder, L. Tan, J. Wang, Y. Liang, N. S. Gray, S. Knapp and A. Chaikuad, *Cell Chem. Biol.*, **2020**, *27*, 1285-1295.e4.
- [155] M. Potowski, V. B. K. Kunig, L. Eberlein, A. Vakalopoulos, S. M. Kast and A. Brunschweiler, *Angew. Chem. Int. Ed.*, **2021**, *60*, 19744–19749.
- [156] A. Sparkes, W. Aubrey, E. Byrne, A. Clare, M. N. Khan, M. Liakata, M. Markham, J. Rowland, L. N. Soldatova, K. E. Whelan, M. Young and R. D. King, *Autom. Exp.*, **2010**, *2*, 1.
- [157] S. O'Neill, *Engineering*, **2021**, *7*, 1351–1353.
- [158] B. Burger, P. M. Maffettone, V. V. Gusev, C. M. Aitchison, Y. Bai, X. Wang, X. Li, B. M. Alston, B. Li, R. Clowes, N. Rankin, B. Harris, R. S. Sprick and A. I. Cooper, *Nature*, **2020**, *583*, 237–241.
- [159] L. Hawizy, D. M. Jessop, N. Adams and P. Murray-Rust, *J. Cheminform.*, **2011**, *3*, 17.
- [160] D. M. Lowe and R. A. Sayle, *J. Cheminform.*, **2015**, *7*, S5.
- [161] A. C. Vaucher, F. Zipoli, J. Geluykens, V. H. Nair, P. Schwaller and T. Laino, *Nat Commun*, **2020**, *11*, 3601.
- [162] Open Reaction Database; <https://open-reaction-database.org/> (accessed on 31.05.2022)
- [163] L. Wilbraham, S. H. M. Mehr and L. Cronin, *Acc. Chem. Res.*, **2021**, *54*, 253–262.
- [164] J. M. Köhler, T. Henkel, A. Grodrian, T. Kirner, M. Roth, K. Martin and J. Metze, *Chem. Eng. J.*, **2004**, *101*, 201–216.
- [165] D. Angelone, A. J. S. Hammer, S. Rohrbach, S. Krambeck, J. M. Granda, J. Wolf, S. Zalesskiy, G. Chisholm and L. Cronin, *Nat. Chem.*, **2021**, *13*, 63–69.
- [166] <https://enamine.net/compound-collections/real-compounds/real-database> (accessed on 11.06.2022)

13 *Appendix*

KNIME Report I - manually scored mediators

Knime report powered by Birt

"column1"	"Min*(Score)"
((R)-2,2'-bis(3,5-tBu-2-HO-C6H2CH=N)-1,1'-binaphthyl)AlCl]	4
((S)-(PhCH(Me))N=C(Me)(Rp-OH[2,2-paracyclophane])	2
((S)-(PhCH(Me))N=C(Phe)(Rp-OH[2,2-paracyclophane])	2
(+)-(1S)-camphor-10-sulphonic acid	4
(+)-(3,2,10-eta-pinene)palladium(II) chloride	0
(+)-1-t-Bu-2-TsO-2H-[1,2]azaboroly*(CO)2*Me3Si*iron	1
(+)-diisopinocampheylboron triflate	4
(+/-)-MIB	3
(-)-1,2-bis((2R,5R)-2,5-dimethylphospholano)benzene	3
(-)-MIB	3
(-)-N-methylephedrine	2
(-)-diisopinocampheylboron triflate	4
(-)-sparteine	2
(1,1'-bis(diphenylphosphino)ferrocene)palladium(II) dichloride	0
(1,2-bis(diphenylphosphanyl)ethane)dichloridopalladium(II)	0
(1,2-dimethoxyethane)dichloronickel(II)	4
(1,3-dimesitylimidazol-2-ylidene)gold(I) chloride	4
(10,10-dimethyl-5-(pyridin-2-yl)-6-azatricyclo[7.1.1.0 ^{2,7}]undeca-2(7),3,5-trien-8-yl)diphenylmethanol	3
(11aR)-(+)-10,11,12,13-tetrahydrodiindeno[7,1-de:1',7'-fg][1,3,2]dioxaphosphocin-5-bis[(R)-1-phenylethyl]amine	2
(11bR)-2,6-di-9-phenanthrenyl-4-hydroxy-dinaphtho[2,1-d:1',2'-f][1,3,2]-dioxaphosphopin-4-oxide	2
(1R)-1-[(1R)-1-[bis[3,5-bis(trifluoromethyl)-phenyl]phosphino]ethyl]-2-[2-bis(4-methoxy-3,5-dimethyl-phenyl)phosphino]-phenyl]ferrocene	2
(1R)-3-di-(3,5-dimethylphenyl)phosphino-(4-diphenylphosphino-2,5-dimethylthienyl-3)-1,7,7-trimethylbicyclo[2.2.1]heptene-2	3
(1R,2R)-2-[(diphenylphosphoroso)amino]-1,2-diphenylethyl]propan-2-yl]amine	2
(1R,2S)-(+)-2-(N,N-di-n-butylamino)-1-phenylpropan-1-ol	2
(1R,2S)-(-)-2-[N-(3,5-di-tert-butylsalicylidene)amino]-1,2-diphenylethanol	2
(1R,2S)-1-phenyl-2-(1-pyrrolidiny)-1-propanol	3
(1R,2S)-2-(4-methylbenzenesulfonylamino)-1,3-diphenyl-1-propanol	3
(1R,2S)-2-Amino-1,2-diphenylethanol	2
(1S)-1-((S)-1-methylpropyl)-(2-morpholin-4-ylethyl)amine	4
(1S)-10-camphorsulfonic acid	4
(1S,2R)-(+)-N-methylephedrine	2
(1S,2R)-(-)-2-(N,N-di-n-butylamino)-1-phenylpropan-1-ol	4
(1S,2R)-(-)-2-(N,N-di-n-propylamino)-1-phenylpropan-1-ol	4

"column1"	"Min*(Score)"
(1S,2S)-2-N,N-dimethylamino-1-(p-nitrophenyl)-3-(tert-butyl)dimethylsilyloxypropan-1-ol	2
(1S,2S,4S,5S)-N2,N2,N5,N5-tetramethylbicyclo[2.2.1]heptane-2,5-diamine	2
(2,7-dimethyl-1,8-biphenylenedioxy)bis(dimethoxyaluminum)	2
(2-hydroxy-ethyl)ammonium acetate	3
(2-methylpropyl)lithium	0
(2R)-(+)-3,3'-diphenyl-[2,2'-dinaphthalene]-1,1'-diol	2
(2R)-(+)-3-exo-N-morpholinoisoborneol	4
(2R,3R)-1,4-dioxaspiro[4.5]decane-alpha,alpha,alpha',alpha'-tetrakis(1-naphthyl)-2,3-dimethanol	4
(2R,3R)-4-dimethylamino-3-methyl-1,2-diphenyl-butan-2-ol	2
(2R,3R)-tartaric acid-derived bis-benzimidazole	4
(2R,3S,3aS,4aR,6R,8aS)-2-isopropyl-6,9,9-trimethyl-3-phenyldecahydro-4aH-pyrrolo[2,1-b][1,3]benzoxazin-3-ol	2
(2S)-(-)-3,3'-diphenyl-(2,2'-binaphthalene)-1,1'-diol	3
(2S)-(?)-3-exo-(morpholino)isoborneol	4
(2S)-2-{diphenyl[(trimethylsilyl)oxy]methyl}pyrrolidine	3
(2S)-N-(2-pyrrolidine-2-carbonyl)-benzenesulfonamide	3
(2S,3S)-2,3-bis(diphenylphosphino)butane	3
(2S,5R)-2-(methylaminomethyl)-1-methyl-5-phenylpyrrolidine	3
(3,5-dioxa-4-phospha-cyclohepta[2,1-a;3,4-a']di-naphthalen-4-yl)-bis(1-phenyl-ethyl)-amine	2
(3S)-(+)-2,2'-diphenyl-(3,3'-biphenanthrene)-4,4'-diol	2
(3aR)-1-methyl-3,3-diphenyl-tetrahydro-pyrrolo[1,2-c][1,3,2]oxazaborole	2
(4R,4'R)-2,2'-(propane-2,2'diyl)bis(4-phenyl-4,5-dihydrooxazole)	2
(4R,5R)-2,2-dimethyl-alpha,alpha,alpha',alpha'-tetra(naphthalen-1-yl)-1,3-dioxolane-4,5-dimethanol	4
(4R,5R)-2-bromo-1,3-bis[(4-methylphenyl)sulfonyl]-4,5-diphenyl-1,3,2-diazaborolidine	4
(4R,5R)-Ph2-1,3-Me2-2-oxo-2-(CH2)5N-1,3,2-diazaphospholidine	2
(4S)-Bn-3-(4-FPhSO2)-2-PhCH2CH2-[1,3,2]-oxazaborolidin-5-one	3
(4S,4'S)-2,2'-(3,6-diphenyl-9H-carbazole-1,8-diyl)bis(4-methyl-4,5-dihydrooxazole)	2
(4S,4'S)-2,2'-(4-chloropyridine-2,6-diyl)bis(4-tert-butyl-4,5-dihydrooxazole)	4
(4S,4S')-(-)-2,2'-(1-methylethylidene)bis[4,5-dihydro-4-(phenylmethyl)oxazole]	2
(4S,5S)-1,3-dimethyl-4,5-diphenyl-2-(1-piperidinyl)-1,3,2-diazaphospholidine 2-oxide	0
(4S,5S)-2-bromo-1,3-bis[(4-methylphenyl)sulfonyl]-4,5-diphenyl-1,3,2-diazaborolidine	4
(5S)-5-benzyl-2,2,3-trimethylimidazolidin-4-one trifluoroacetic acid salt	4
(5aR,10bS)-2-mesityl-5a,10b-dihydro-4H,6H-indeno[2,1-b][1,2,4]triazolo[4,3-d][1,4]oxazin-2-ium tetrafluoroborate	1

"column1"	"Min*(Score)"
(5aS,10bR)-5a,10b-dihydro-2-(2,4,6-trimethylphenyl)-4H,6H-indeno[2,1-b]-1,2,4-triazolo[4,3-d]-1,4-oxazinium chloride	2
(6,8,15,17-tetramethyldibenzo[b,i][1,4,8,11]-tetraazacyclo-tetradecinato)nickel(II)	4
(C2H5)2AlSnClF2	0
(E)-1,1-bis(3,5-dimethylphenyl)-N-(pyridin-2-yl-methylene)methanamine iron(II) chloride dichloromethane solvate	4
(E)-diazene-1,2-diylbis(cyclohexylmethanone)	2
(IPr)Au(NTf2)	0
(R)-((4,4?-bi-1,3-benzodioxole)-5,5?-diyl)bis(bis(3,5-di-t-butyl-4-methoxyphenyl))phosphine	2
(R)-(-)-3,3?-bis(3,5-diphenylphenyl)-1,1?-binaphthalene-2,2?-sulfonimide	3
(R)-(3,3'-bis(1-naphthyl)-1,1'-binaphthanele-2,2'-yl)phosphoric acid	3
(R)-(?)-1-[(R)-2-(2?-diphenylphosphinophenyl)ferrocenyl]ethylbis(di-3,5-trifluoromethylphenyl)phosphine	3
(R)-1,1'-Bi-2-naphthol	1
(R)-1,1'-binaphthalene-2,2'-diol lithium salt	2
(R)-1,1'-binaphthyl-2,2'-phosphoric acid	4
(R)-1-{(RFc)-2-[2-(diphenylphosphino)phenyl]ferrocenyl}ethylbis[3,5-bis-(trifluoromethyl)phenyl]phosphine	3
(R)-10-camphorsulfonic acid	4
(R)-2,2'-bis(diphenylphosphanyl)-1,1'-binaphthyl	4
(R)-2,2'-bis[bis(3-methylphenyl)phosphino]-1,1'-binaphthyl	4
(R)-2,2'-dihydroxy-1,1'-binaphthyl	4
(R)-2,2?-diphenyl-(4-biphenanthrol)	4
(R)-2,6-bis(naphthalen-2-yl)-4-oxo-3,5-dioxa-4lambda5-phosphacyclohepta[2,1-a;3,4-a']dinaphthalen-4-ol	2
(R)-2-(diphenyl(trimethylsilyloxy)methyl)pyrrolidine	3
(R)-2-methyl-1-(2-(3,3-dimethylbut-1-ynyl)pyrimidin-5-yl)propan-1-ol	3
(R)-3,3',6,6'-tetraiodo-1,1'-binaphthalene-2,2'-diol	3
(R)-3,3'-bis(2,4,6-triisopropylphenyl)-1,1'-binaphthyl-2,2'-diylhydrogenphosphate	4
(R)-3,3'-bis(2,4,6-triisopropylphenyl)binol phosphoric acid	4
(R)-3,3'-bis(4-trifluoromethylphenyl)-1,1'-binaphthyl-2,2'-diylphosphoric acid	4
(R)-3,3'-bis(9-anthracenyl)-1,1'-binaphthyl-2,2'-diyl hydrogenphosphate	4
(R)-3,3'-bis(dimethylphenylsilyl)-1,1'-binaphthyl-2,2'-dicarboxylic acid	4
(R)-3,3'-bis(triphenylsilyl)-1,1'-bi-naphthyl-2,2'-diyl	4
(R)-3,3'-di(anthracen-9-yl)-5,5',6,6',7,7',8,8'-octahydro-[1,1'-binaphthalene]-2,2'-diyl hydrogen phosphate	3
(R)-3,3'-dibromo-1,1'-bi-2-naphthol	1
(R)-3,3'-dichloro-1,1'-binaphthalene-2,2'-diol lithium salt	2
(R)-3,3'-diiodo-2,2'-dihydroxy-1,1'-binaphthyl	4

"column1"	"Min*(Score)"
(R)-3,3'-bis(2,4,6-triisopropylphenyl)-BINOL-phosphoric acid	0
(R)-3,3'-difluoro-1,1'-bi-2-naphthol	1
(R)-3-(3,5-diphenylphenyl)-2,2'-dihydroxy-5,5',6,6',7,7',8,8'-octahydro-1,1'-binaphthyl	4
(R)-3-diphenylhydroxymethyl-2,2'-dihydroxy-1,1'-bi-naphthalenyl	3
(R)-BINAPHANE	2
(R)-DHTP	3
(R)-SEGPPOS-I	1
(R)-methylaluminum beta-binaphthoxide	0
(R)-p-Tol-BINAP*AgF	0
(R)-segphos	1
(R)?2,2'?bis(diphenylphosphoryl)?1,1'?binaphthyl	4
(R,R')-N,N'-bis(5-tert-butyl-2-hydroxybenzylidene)-1,2-cyclohexanediamine	2
(R,R)-1,2-bis(2,5-diphenylphospholanyl)ethane	3
(R,R)-1,2-diphenylethandiol (2-methoxy)ethyl diether	2
(R,R)-N,N'-dimethylstilbene-1,2-diamine chiral phosphoramidate	2
(R,R)-TADDOL	0
(R,R)-hydroxybenzoin	3
(R,R)-walphos	3
(R,S)-2-OH-3,5-Cl ₂ -C ₆ H ₂ -SO ₂ -NH-CH(CH ₂ Ph)-CH(Ph)OH	2
(Ra)-5,6,7,8,5',6',7',8'-octahydro-[1,1']binaphthalenyl-2,2'-diol	3
(Ra)-N-[(1S,2R)-1,2-diphenyl-2-hydroxyethyl]-3,5-dihydro-4H-dinaphtho[2,1-c:1',2'-e]-azepine	3
(Re(CO) ₃ (THF)Br) ₂	0
(RhCl(diene*)) ₂	0
(S)-(-)-2,2'-dihydroxy-1,1'-binaphthalene	3
(S)-(-)-2,2'-dihydroxy-1,1'-binaphthyl	4
(S)-(1,1'-binaphthalene)-2,2'-diylbis(diphenylphosphine)	3
(S)-1,1'-binaphthalene-2,2'-diylbis(diphenylphosphineoxide)	3
(S)-1-methyl-2-(1-naphthylaminomethyl)pyrrolidine	3
(S)-2'-methoxy-2-methylthio-1,1'-binaphthalene	3
(S)-2,2',5,5'-tetramethyl-4,4'-bis-(diphenylphosphino)-3,3'-bithiophene oxide	2
(S)-2-(((tert-butyl)dimethylsilyloxy)diphenylmethyl)pyrrolidine	3
(S)-2-(1-pyrrolidinylmethyl)pyrrolidine	3
(S)-2-(diphenyl((triethylsilyloxy)methyl)pyrrolidine	3
(S)-2-[bis(3,5-bis(trifluoromethyl)phenyl)-triethyl-siloxy-methyl]-pyrrolidine	3
(S)-3,3'-bis(2,4,6-tri-iso-propylphenyl)-1,1'-bi-naphthyl-2,2'-diyl hydrogenphosphate	4
(S)-3,3'-bis(4"-tert-butylphenyl)-2,2'-(2,2-bisbromo-2-stannopropane-1,3-diyl)-1,1'-binaphthyl	4
(S)-3,3'-diiodo-2,2'-dihydroxy-1,1'-binaphthyl	4
(S)-3,3'-dimethyl-[1, 1'-binaphthalene]-2, 2'-diol	3

"column1"	"Min*(Score)"
(S)-3,3'-bis(9-anthracenyl)-1,1'-binaphthyl-2,2'-diyl N-triflyl-phosphoramidate	4
(S)-3,5-dichloro-N-(6-(4-isopropyl-4,5-dihydrooxazol-2-yl)-2,3-dimethoxyphenyl)benzenesulfonamide	3
(S)-5-benzhydryl-2-(perfluorophenyl)-6,7-dihydro-5H-pyrrolo[2,1-c][1,2,4]triazol-2-ium tetrafluoroborate	1
(S)-5-benzyl-2-(2,6-dimethoxyphenyl)-6,6-dimethyl-6,8-dihydro-5H-[1,2,4]triazolo[3,4-c][1,4]oxazin-2-ium tetrafluoroborate	1
(S)-5-benzyl-2-mesityl-6,6-dimethyl-5,6-dihydro-8H-[1,2,4]triazolo[3,4-c][1,4]oxazin-2-ium tetrafluoroborate	1
(S)-6,6'-bis(2,4,6-triisopropylphenyl)-1,1'-spirobiindane-7,7'-diyl hydrogenphosphate	2
(S)-6,6'-di(naphthalen-1-yl)-1,1'-spirobiindane-7,7'-diyl phosphate	2
(S)-AlCl[2,2'-(O-3,5-(t-Bu)2-C6H2-CH=N)]2-1,1'-binaphthyl]	4
(S)-N-((3-methylpyridin-2-yl)carbamothioyl)pyrrolidine-2-carboxamide	4
(S)-[1,1'-binaphthalen]-2-ylidiphenylphosphine	3
(S)-[1,1']-binaphthalenyl-2,2'-diol	3
(S)-bis(4-fluorophenyl)(1-methylpyrrolidin-2-yl)methanol	3
(S,S)-6,6'-bis(1-hydroxy-2,2-dimethylpropyl)-2,2'-bipyridine	3
(S,R)-N-PINAP	2
(S,S)-(+)-2,6-bis[2-(hydroxydiphenylmethyl)-1-pyrrolidinyl-methyl]-4-methylphenol	3
(S,S)-1,1'-(6,6',7,7'-tetrahydro-5H,5'H-[1,1'-bi(cyclopenta[c]pyridine)]-3,3'-diyl)bis(2,2-dimethylpropan-1-ol)	4
(S,S)-2,2'-methylenebis(4-tert-butyl-2-oxazoline)	2
(S,S)-4-tBu-2,6-bis[2-(HOPh2C-)pyrazolidin-1-ylmethyl]phenol	2
(S,S)-Bn-bod	2
(carbonyl)chloro(hydrido)tris(triphenyl-phosphine)ruthenium(II)	0
(eta6-toluene)Ni(1,3-bis(2,6-diisopropylphenyl)imidazolin-2-ylidene)	4
(Ipc)2BH	2
(methyl benzoate)chromium tricarbonyl	0
(mu3,eta2,eta3,eta5-acenaphthylene)Ru3(CO)7	4
(o,o'-biphenylenedioxy)methylaluminium	0
(oxydi-2,1-phenylene)bis(diphenylphosphine)Pd(pi-allyl)Cl	0
(pi-allyl)palladium chloride	0
(polyallyl)scandium trifylamide ditriflate	4
(tricyclohexylphosphine)gold(I) chloride	4
(triphenyl phosphite)gold(I) chloride	4
(triphenylphosphine)gold(I) chloride	4
(±)N,N'-bis(3,5-di-tert-butylsalicylidene)-1,2-cyclohexanediamine cobalt(II)	0
1,1'-(1,2-ethanediyl)bisbenzene	2
1,1'-bi-2-naphthol	4
1,1'-binaphthyl-2,2'-diyl hydrogenphosphate	4
1,1'-biphenyl-2,2'-diyl hydrogen phosphate	2
1,1'-bis(di-tertbutylphosphino)ferrocene	2

"column1"	"Min*(Score)"
1,1'-bis(dicyclohexylphosphinocyclopentadienyl iron	4
1,1'-bis(diisopropylphosphino)ferrocene	2
1,1'-bis-(diphenylphosphino)ferrocene	3
1,1'-carbonyldiimidazole	3
1,1,1,3',3',3'-hexafluoro-propanol	2
1,1,1,3,3,3-hexamethyl-disilazane	0
1,1,3,3-tetramethylguanidine	4
1,1-dicyclohexyl-N-(dicyclohexylphosphino)-N-methylphosphinamine	2
1,1-dimethoxyethane	2
1,10-Phenanthroline	2
1,1'-bi-2-naphthol	4
1,1'-binaphthalene-2,2'-diylbis[bis(4-methylphenyl) phosphine]	3
1,2,2,6,6-pentamethylpiperidine	4
1,2,3-Benzotriazole	4
1,2-bis(2,5-dimethylphospholano)benzene	3
1,2-bis(dimethylphosphanyl)ethane	2
1,2-bis(diphenylphosphino)ethane nickel(II) chloride	4
1,2-bis-(diphenylphosphino)ethane	3
1,2-bisethane	2
1,2-dichloro-ethane	4
1,2-dimethyl-3-[4-(1,2-dimethyl-1H-imidazol-3-ium-3-yl)butyl]-1H-imidazol-3-ium dibromide	4
1,3,4,6,7,8-hexahydro-2H-pyrimido[1,2-a]pyrimidine	3
1,3-bis(2,4,6-trimethylphenyl)-4,5-dihydro-imidazolium chloride	4
1,3-bis(2,6-(i-Pr) ₂ -phenyl)-4,5-dihydroimidazolin-2-ylidene	4
1,3-bis(2,6-diethylphenyl)-1H-imidazol-3-ium chloride	2
1,3-bis(cyclohexyl)imidazolium tetrafluoroborate	1
1,3-bis(mesityl)imidazolium chloride	4
1,3-bis-(diphenylphosphino)propane	3
1,3-bis[2,6-diisopropylphenyl]imidazolium chloride	4
1,3-dibenzyl-1H-benzo[d]imidazol-3-ium chloride	2
1,3-dimethyl-2-imidazolidinone	4
1,3-dimethyl-3,4,5,6-tetrahydro-2(1H)pyrimidinone	3
1,3-dimethylimidazolim iodide	4
1,4-di(diphenylphosphino)-butane	3
1,4-diaminobutane	4
1,4-diaza-bicyclo[2.2.2]octane	2
1,4-dimethyl-1,2,4-triazolium iodide	4
1,4-phenylenediacetic acid	4
1,5-diazabicyclo[5.4.0]-undec-7-ene	4
1,5-diazabicyclo[5.4.0]undecene	4

"column1"	"Min*(Score)"
1,8-diazabicyclo[5.4.0]undec-7-ene	4
1-(2,6-diisopropylphenyl)-3-(2-(phenylthio)phenyl)-4,5-dihydroimidazolium chloride	4
1-(3,5-Bis-trifluoromethyl-phenyl)-3-((1R,2R)-2-dimethylamino-cyclohexyl)-thiourea	3
1-(3,5-Bis-trifluoromethyl-phenyl)-3-[(S)-quinolin-4-yl-((2R,4S,5R)-5-vinyl-1-aza-bicyclo[2.2.2]oct-2-yl)-methyl]-thiourea	3
1-(3,5-bis(trifluoromethyl)phenyl)-3-((1R)-(6-meth-oxyquinolin-4-yl)(3-vinylquinuclidin-7-yl)methyl)urea	2
1-(3,5-bis(trifluoromethyl)phenyl)-3-((1R,2R)-2-(piperidin-1-yl)cyclohexyl)thiourea	3
1-(3,5-bis(trifluoromethyl)phenyl)-3-((1R,2R)-2-(pyrrolidin-1-yl)cyclohexyl)thiourea	3
1-(3,5-bis(trifluoromethyl)phenyl)-3-((1S)-(6-hydroxyquinolin-4-yl)(5R)-5-vinylquinuclidin-2-yl)methyl)thiourea	3
1-(3,5-bis(trifluoromethyl)phenyl)-3-((1S)-(6-meth-oxyquinolin-4-yl)(5-vinylquinuclidin-2-yl)methyl)thiourea	2
1-(3,5-bis(trifluoromethyl)phenyl)-3-((S)-(6-meth-oxyquinolin-4-yl)((2S,4S,8R)-8-vinylquinuclidin-2-yl)methyl)thiourea	2
1-(phenylsulfonyl)propyne	2
1-(tert-butoxycarbonyl)-L-proline	4
1-[3,5-bis(trifluoromethyl)phenyl]-3-[(1S,2S)-2-(dimethylamino)cyclohexyl]thiourea	3
1-[3,5-bis(trifluoromethyl)phenyl]-3-phenyl-2-thiourea	3
1-[bis(trifluoromethanesulfonyl)methyl]-2,3,4,5,6-pentafluorobenzene	2
1-acetoxy-1,2-benziodoxol-3-one	3
1-butyl-3-methylimidazolium Tetrafluoroborate	0
1-butyl-3-methylimidazolium hydroxide	0
1-fluoro-2,4,6-trimethylpyridin-1-ium tetrafluoroborate	1
1-hydrosilatrane	2
1-hydroxy-3H-benz[d][1,2]iodoxole-1,3-dione	2
1-methoxy-2-methyl-1-trimethylsiloxy-1-propene	2
1-methyl-1H-imidazole	4
1-methyl-2,3,4,6,7,8-hexahydro-1H-pyrimido[1,2-a]pyrimidine	3
1-methyl-3-(2,4,6-trimethylphenyl)-3H-benz-imidazol-1-ium iodide	4
1-methyl-3-(4-sulfobutyl)-1H-imidazol-3-ium hydrogensulfate	4
1-methyl-3-methylimidazol-3-ium dimethyl phosphate	2
1-methyl-piperazine	2
1-methyl-pyrrolidin-2-one	3
1-methylimidazole-3-sulfonic acid hydrochloride	4
1-n-butyl-3-methylimidazolim bromide	4
1-naphthalenesulfonic acid	0
1-pyrroline	3
1-{3,5-bis(trifluoromethyl)phenyl}-3-{(1R,2R)-2-(dimethylamino)cyclohexyl}thiourea	3
1-{3,5-bis(trifluoromethyl)phenyl}-3-{(1R,2R)-2-(pyrrolidin-1-yl)cyclohexyl}urea	3

"column1"	"Min*(Score)"
10 percent chiral ammonium fluoride	0
10 wt% Pd(OH) ₂ on carbon	1
10-camphorsulfonic acid	4
10-camphorsulfonic acid	4
10-methyl-9-(2,4,6-trimethylphenyl) acridinium tetrafluoroborate	1
10V/SiO ₂ -25	4
15-crown-5	1
18-crown-6 ether	1
18O-labeled water	4
1H-imidazole	4
2,2'-bis(diphenylphosphino)-5,5',6,6',7,7',8,8'-octahydro-1,1'-binaphthyl	4
2,2'-bis-(diphenylphosphino)-1,1'-binaphthyl	4
2,2,2-trifluoroethanol	4
2,2,6,6-Tetramethyl-1-piperidinyloxy free radical	0
2,2,6,6-tetramethyl-piperidine	4
2,2,6,6-tetramethyl-piperidine-N-oxyl	0
2,2,6,6-tetramethylpiperidinyllithium	0
2,2,6,6-tetramethylpiperidinylmagnesium chloride	4
2,2,6,6-tetramethylpiperidinylmagnesium chloride lithium chloride complex	4
2,2-dimethylthiolane	2
2,2'-azobis(4-methoxy-2,4-dimethyl)valeronitrile	4
2,2'-methylene bis[(4R,5S)-4,5-diphenyl-2-oxazoline]	2
2,3,4-trimethoxy-N-((R)-quinolin-4-yl)((1S,2S,4S,5R)-5-vinylquinuclidin-2-yl)methylbenzenesulfonamide	3
2,3-dicyano-5,6-dichloro-p-benzoquinone	0
2,3-dihydro-1H-1,3-dimesylimidazole-2-carbene	4
2,4,6-trimethyl-pyridine	4
2,4,6-tripropyl-1,3,5,2,4,6-trioxatriphosphinane-2,4,6-trioxide	0
2,4-dinitrobenzoic acid	3
2,5-dimethyl-piperazine	2
2,6-bis((R)-4-phenyl-4,5-dihydrooxazol-2-yl)pyridine	4
2,6-bis(bis(3,5-bis(trifluoromethyl)phenyl)(hydroxy)-methyl)-dinaphtho-[2,1-d:1',2'-f][1,3,2]dithiazepine 3,3,5,5-tetraoxide	3
2,6-bis(hydroxybis(3-(trifluoromethyl)phenyl)methyl)dinaphtho[2,1-d:1',2'-f][1,3,2]dithiazepine 3,3,5,5-tetraoxide	3
2,6-bis(pyrazole)pyridine	4
2,6-bis-(4-chloro-phenyl)-4-oxo-3,5-dioxo-4-lambda5-phosphacyclohepta[2,1-a;3,4-a']di-naphthalen-4-ol	2
2,6-bis<5',5'-diphenyl-4'-(S)-isopropyl oxazolin-2'-yl>pyridine	4
2,6-bis[(R,R)-4-(1-TPSO-ethyl)-2-oxazolin-2-yl]pyridine	4
2,6-bis[4'-(S)-(tert-butyl)oxazolin-2'-yl]pyridine	4
2,6-di-tert-butyl-4-methylpyridine	4
2,6-di-tert-butyl-pyridine	4
2,6-dimethylpyridine	4

"column1"	"Min*(Score)"
2,7-dimethyl-1,8-biphenylenediol	2
2-(((1R,2R)-2-(3-(3,5-bis(trifluoromethyl)phenyl)thioureido)cyclohexyl)carbamoyl)-3,4,5,6-tetrabromobenzoic acid	3
2-(((2,6-diisopropylphenyl)imino)methyl)pyridine	4
2-(2-hydroxyethyl)-3,4-dimethylthiazolium iodide	4
2-(3-(cyclopropylmethoxy)phenoxy)-4-fluorobenzaldehyde	3
2-(4-bromophenyl)-acetic acid	4
2-(5,5-dimethyl-1,3,2-dioxaborinan-2-yl)-5,5-dimethyl-1,3,2-dioxaborinane	2
2-(N,N-dimethylamino)athanol	2
2-(aminomethylcyclohexyl)ethylamine	2
2-(di-tert-butylphosphino)-1,1'-biphenylgold(I) chloride	4
2-(diphenylphosphino)-N-((S)-((1S,2S,4S,5R)-5-ethyl-quinuclidin-2-yl)(6-methoxyquinolin-4-yl)methyl)benzamide	3
2-(tert-butylethynyl)pyrimidine-5-carbaldehyde	3
2-(trimethylsilyl)phenyl trifluoromethanesulfonate	2
2-Methylpiperidin	4
2-amino-2-hydroxymethyl-1,3-propanediol	3
2-aminopyridine	4
2-chloropyridine	4
2-cyano-2-(hydroxyimino)acetic acid methylester	3
2-fluoro-2-iodo-1,3-benzodithiole-1,1,3,3-tetraoxide	2
2-hydroxy-p-toluic acid	1
2-hydroxyethanethiol	2
2-iodoxybenzoic acid	4
2-isopropoxy-4,4,5,5-tetramethyl-2-vinyl-1,3,2-dioxasilolane	2
2-mesityl-2,5,6,7-tetrahydropyrrolo[2,1-c][1,2,4]triazolium chloride	4
2-mesityl-6,7-dihydro-5H-pyrrolo[2,1-c][1,2,4]triazol-2-ium chloride	2
2-mesityl-6,7-dihydro-5H-pyrrolo[2,1-c][1,2,4]triazol-2-ium tetrafluoroborate	1
2-methyl-but-2-ene	4
2-nitropropane	4
2-pentafluorophenyl-6,7-dihydro-5H-pyrrolo[2,1-c][1,2,4]triazol-2-ium tetrafluoroborate	1
2-phenyl-6,7-dihydro-5H-pyrrolo[2,1-c][1,2,4]triazol-2-ium tetrafluoroborate	1
2-tert-butylanthraquinine	0
2.9-dimethyl-1,10-phenanthroline	1
20% palladium hydroxide-activated charcoal	0
20percent iron-modified mesoporous silica SBA-15	4
2ClO4(1-)*7.25H2O*Fe(2+)	4
2V/SiO2-500	4
3 A molecular sieve	0
3 Angstroem MS	0

"column1"	"Min*(Score)"
3,3'-(1,4-phenylenebis(methylene))bis(5-(2-hydroxyethyl)-4-methylthiazol-3-ium) bromide	3
3,3'-bis-pyrrolidin-1-ylmethyl-[1,1']binaphthalenyl-2,2'-diol	2
3,3'-di(9-phenylanthryl)BINOL phosphoric acid	0
3,3'-dibromo-2,2'-dihydroxy-1,1'-binaphthyl	4
3,4,5,6-tetrahydropyrimidin-2(1H)-one	3
3,4-dimethyl-2-(alpha-hydroxybenzyl)thiazoliumiodide	4
3,4-dimethyl-5-(2-hydroxyethyl)thiazolium iodide	4
3,5,3',5'-tetra-tert-butyl-4,4'-diphenoquinone	0
3,5-difluoropyridine	4
3,6-di(2'-pyridyl)-1,2,4,5-tetrazine	2
3,6-di-tert-butyl-9,10-dimesitylacridinium tetrafluoroborate	1
3,6-dioxocyclohexa-1,4-diene-1,2-dicarbonitrile	0
3-((3,5-bis(trifluoromethyl)benzyl)amino)-4-(((1S)-(6-methoxyquinolin-4-yl)((2S,4S,5R)-5-vinylquinuclidin-2-yl)methyl)amino)cyclobut-3-ene-1,2-dione	3
3-Dimethylamino-1-propanol	2
3-Methyl-1-phenyl-2-phospholene 1-oxide	0
3-Methylpiperidine	4
3-amino propanoic acid	4
3-azapentane-1,5-diamine	1
3-benzyl-4,5-dimethylthiazol-3-ium bromide	3
3-benzyl-5-(2-hydroxyethyl)-4-methyl-1,3-thiazol-3-ium chloride	3
3-chloro-benzenecarboxylic acid	2
3-ethyl-5-(2-hydroxyethyl)-4-methyl-1,3-thiazolium bromide	4
3-ethyl-5-(2-hydroxymethyl)-4-methyl-1,3-thiazolium bromide	4
3-iodo-2-methylcyclohexenone	2
3-nitrobenzoic acid	4
4 A molecular sieve	0
4 A molecular sieves	0
4 Angstroem M.S	2
4 Angstroem MS	0
4 Angstroems MS	0
4'-(4-methylphenyl)-2,2':6,2'-terpyridine	4
4,4'-di-tert-butyl-2,2'-bipyridine	4
4,4'-di-tert-butylbiphenyl	2
4,4'-dibromobpy-NiCl2	4
4,5-bis(diphenylphos4,5-bis(diphenylphosphino)-9,9-dimethylxanthenephino)-9,9-dimethylxanthene	3
4-[(5S)-3-ethyl-4-oxa-1-azatricyclo[4.4.0.0 ^{3,8}]decan-5-yl]quinolin-6-yl (2,2,2-trichloroacetyl)carbamate	3
4-fluorobenzyl alcohol	4
4-hydroxy-2,6-di(naphthalen-2-yl)dinaphtho[2,1-d:1',2'-f][1,3,2]dioxaphosphepine 4-oxide	2
4-methoxy-2,2,6,6-tetramethylpiperidin-1-oxyl radical	0
4-methoxy-N-(pyridin-4-ylmethylene)benzenamine	3

"column1"	"Min*(Score)"
4-methoxy-N-[(E)-phenylmethylidene]aniline	4
4-methoxy-aniline	4
4-methoxy-phenol	3
4-methyl-morpholine	4
4-methylpiperidin	4
4-morpholineethanesulfonic acid	4
4-nitraminopyridine N-oxide	4
4-nitro-benzoic acid	4
4-toluenesulfonyl azide	0
4A MS	0
5 wtpercentFe-H-Beta-SiO2/Al2O3=150 zeolite	4
5%-palladium/activated carbon	1
5,10,15,20-tetrakis(p-chlorophenyl)porphyrin iron(III) chloride	4
5,5'-dimethyl-2,2'-bipyridine	4
5,5'-bis[di(3,5-di-tert-butyl-4-methoxyphenyl)phosphino]-4,4'-bi-1,3-benzodioxole	2
5-(2-hydroxy-ethyl)-3,4-dimethyl-thiazolium chloride	4
5-ethyl-2-methylpyridine borane complex	4
5-iodo-2,3-dihydrobenzofuran	2
5-methoxy-1H-benzimidazole	2
5A molecular sieve	0
5a(S),10b(R)-5a,10b-dihydro-2-(pentafluorophenyl)-4H,6H-indeno[2,1-b][1,2,4]triazolo[4,3-d][1,4]oxazinium tetrafluoroborate	1
5percent iron-modified mesoporous silica SBA-15	4
7b-phenyl-7bH-oxazirino<2,3-b><1,2>benz-isothiazole 3,3-dioxide	2
9,10-phenanthrenequinone	0
9-BBN triflate	4
9-benzylfluorene lithium salt	2
AD-mix-beta	2
ATP	4
Adam?s catalyst	0
Ag*ClO4(1-)=AgClO4(1-)	4
AlPO4 supported ethylenediamine-chromium(III)-salen complex nanoparticles	0
Amberlite G-50 ion-exchange resin	1
Amberlite IRA-400	1
Amberlyst 15	1
Amberlyst A-21 ion-exchange resin	1
Ambersep 900 OH resin	1
Au(OAc)3	0
B-(trifluoromethanesulfonyloxy)-9-borabicyclo[3.3.1]nonane	3
BCl(C6H11)2	2
BF3*(OEt)2	2

"column1"	"Min*(Score)"
BF4(1-)*2C8H12*Rh(1+)	0
BF4(1-)*C18H13Cl3N3O(1+)	2
BF4(1-)*C18H17F5N3O(1+)	2
BF4(1-)*C21H22N3O(1+)	2
BF4(1-)*Cu(2+)*6H2O	2
BIPHEP-I	2
Bis<2-(N,N-dimethylamino)ethyl>aether	2
Bromotrichloromethane	4
Bu3SnN(Ph)C(OMe)=NPh	0
BuLi	0
C42H63AuO3P(1+)*C2F6NO4S2(1-)	0
C49H68AuNP(1+)*F6Sb(1-)	0
C5H5BrMnO5	0
CH3BN(1-)*C10H15NPol(1+)	0
CH3BN(1-)*H(1+)*H2NPol	0
Candida rugosa lipase	2
Carbonate buffer	4
Celite	4
Cinchonin	2
Cl(3)HO4	2
Co(3,5-DitBu-Ibu-Phyrin)	0
Co(meso-tetraphenylporphyrin) tetrakis[3,5-bis(trifluoromethyl)phenyl]borate	0
Coenzyme A	4
CrCl2*(THF)1.4	0
CrCl2*DMF	0
CuCl*2LiCl	0
CuF(PPh3)3 methanol solvate	0
CuF*(R)-tolBinap	0
CuF-(R)-DTBM-SEGPPOS	0
CuI*2LiBr	0
CyJohnPhos	2
D-Prolin	4
DBN	2
DIMCARB	2
DL-dithiothreitol	2
DTBP	2
Dess-Martin periodane	1
Diethoxy-methyl-(6-amino-hexylaminomethyl)-silan	2
Difluoroacetic acid	4
Diphenyl(N-methyl-2-pyrrolidiny)methanol	3

"column1"	"Min*(Score)"
Eaton's reagent	0
Echavarren's catalyst	4
Ethyl diphenylphosphinite	3
Ethyl propionate	3
Fe(TCP)Cl	4
GeCl ₂ *dioxane	0
GeCl ₂ -dioxane	0
Grotjahn's catalyst	4
H ₂ SO ₄ -SiO ₂	0
HATU	4
Hexamethylbenzene	0
Hexamethyldisiloxane	2
Hexamethylphosphorous triamide	0
Hf(OTf) ₄	0
Hoveyda-Grubbs catalyst second generation	3
In(OSO ₂ CF ₃) ₃	0
Indion-130 resin	4
Iron(III) nitrate nonahydrate	4
Isopropyl acetate	4
K10 montmorillonite clay	4
L-(+)-diisopropyl tartrate	4
L-Cysteine	4
L-Leucine supported on superparamagnetic silica encapsulated gamma-Fe ₂ O ₃ nanoparticles	4
L-Tartaric acid	4
L-diisopinocampheylborane	0
L-proline	4
L-prolinium sulfate	4
Lawessons reagent	1
Lewatit S 100 ion exchange resin	4
Li(2,2,6,6-tetramethylpiperidide)*Al(iBu) ₃	0
LiHMDSA	0
Lindlar's catalyst	4
MANDELIC ACID	0
MIL-101-SO ₃ H	2
MP-BH(OAc) ₃ resin	4
MP-BH ₃ CN	2
MP-CNBH ₃	2
MP-cyanoborohydride	0
MP-triacetoxyborohydride	0
MP-triacetoxyborohydride resin	4

"column1"	"Min*(Score)"
MS 4 Angstroem	0
Mesitol	4
Methyltrichlorosilane	0
Mg-Al hydrotalcite	2
Mn89Cr11	0
MnBr(CO)5	0
Mo(CO)3(CN-t-Bu)3	1
Montmorillonite K 10	4
Montmorillonite K10 clay	4
Montmorillonite KSF	0
N+C5Ala2C16	2
N,N'-((11bS,11b'S)-azanediylbis(2,6-bis(3,5-bis(pentafluoro-lambda6-sulfanyl)phenyl)-4lambda5-dinaphtho[2,1-d:1',2'-f][1,3,2]dioxaphosphepine-4-yl-4-ylidene))bis(1,1,2,2-pentafluoroethane-1-sulfonamide)	2
N,N'-((11bS,11b'S)-azanediylbis(2,6-bis(3,5-bis(perfluoropropyl)phenyl)-4lambda5-dinaphtho[2,1-d:1',2'-f][1,3,2]dioxaphosphepine-4-yl-4-ylidene))bis(1,1,1-trifluoromethanesulfonamide)	0
N,N'-Dimethylurea	3
N,N'-dimethylpiperazine	2
N,N,N',N'',N'''-pentamethyldiethylenetriamine	2
N,N,N',N'-Tetraethylethylenediamine	4
N,N,N',N'-tetramethyl-1,4-butanediamine	2
N,N,N',N'-tetramethyl-1,8-diaminonaphthalene	4
N,N,N',N'-tetramethyl-1,8-diaminonaphthalene	4
N,N,N',N'-tetramethylguanidine	4
N,N,N,N,-tetramethylethylenediamine	4
N,N,N,N,N,N-hexamethylphosphoric triamide	1
N,N,N-triethyl-N-(propanesulfonic acid)ammonium hydrogensulfate	0
N,N,N?,N?-tetramethyl-N?-tert-butylguanidine	4
N,N-di(propan-2-yl)-4H-1,3,2-benzo-dioxaborinin-2-amine	2
N,N-dibutyl amino-2 ethanol	1
N,N-diisopropyl-1,2-ethanediamine	2
N,N-dimethyl acetamide	4
N,N-dimethyl-aniline	4
N,N-dimethyl-ethanamine	2
N,N-dimethyl-formamide	4
N,N-dimethylalanine	4
N,N-dimethylammonium chloride	1
N,N-dimethylethylenediamine	4
N,N?-bis(2,6-diisopropylphenyl)imidazol-2-ylidene hydrochloride	4
N,N`-dimethylethylenediamine	4
N,O-bis-(trimethylsilyl)-acetamide	4
N-(2-acetamido)-3-iminodiacetic acid	4

"column1"	"Min*(Score)"
N-(3,5-bis(trifluoromethyl)phenyl)-3-((2-(((S)-6-methoxy-quinolin-4-yl)((1S,2S,4S,5R)-5-vinylquinuclidin-2-yl)methyl)amino)-3,4-dioxocyclobut-1-en-1-yl)amino)-5-(trifluoromethyl)benzamide	3

"column1"	"Min*(Score)"
N-(p-toluenesulfonyl)-L-valine	4
N-(tert-butyl)benzenesulfinimidoyl chloride	2
N-3,5-(CF3)2-C6H3-N'-[9-dehydroxy-quinidin-9(S)-yl]thiourea	3
N-Bromosuccinimide	4
N-[3,5-bis(trifluoromethyl)phenyl]-N'-[(9R)-6'-methoxycinchonan-9-yl]thiourea	2
N-benzyl-N,N,N-triethylammonium chloride	1
N-benzyl-trimethylammonium hydroxide	0
N-benzylidenephenylylsulfonamide	2
N-butylamine	2
N-chloro-succinimide	4
N-cyclohexyl-cyclohexanamine	2
N-ethyl-N,N-diisopropylamine	3
N-ethylmorpholine	4
N-fluorobis(benzenesulfon)imide	2
N-iodo-succinimide	4
NAD	4
Na(1+)*HSO4(1-)*SiO2 = NaHSO4*SiO2	0
Na(OAc)3BH loaded resin	1
Na2H2S2O5	0
Ni(acetylaceta)2	4
NiCl(o-tolyl)(tetramethylethylenediamine)	4
Noyori's catalyst	4
O,O-Diethyl hydrogen phosphorodithioate	0
O-(1H-benzotriazol-1-yl)-N,N,N',N'-tetramethyl-uronium hexafluorophosphate	1
O-(4-nitrobenzoyl) hydroxylamine	0
O4S(2-)*2Al(3+)*4CH3(1-)	2
Oxone	0
P(p-C6H4F)3	2
P(p-CH3OC6H4)3	2
PL-cyanoborohydride resin	4
PS-CNBH3	2
PS-Trisamine	0
PS-cyanoborohydride	0
PS-diisopropylethylamine	4
PS-isocyanate scavenger resin	4
PS-p-toluensulfonyl hydrazide scavenger resin	4
PS-triacetoxyborohydride	0
PYRIMIDINE	3
PdCl(dppb)(C3H5)	0
Pentafluorobenzoic acid	4
Pic-BH3	2

"column1"	"Min*(Score)"
Quinuclidine	0
Rh2(OAc)4	0
Rh2(esp)2	0
Rh2(trifluoroacetate)4(1,3-bis(2,6-diisopropylphenyl)imidazol-2-ylidene)2	0
Rh2[3S-3-(1,3-dioxobenzo[f]isoindol-2-yl)-2-piperidinonate]4	0
RhHCl2(PPh3)3	0
Ru(2 wt%)/CeO2	4
Ru2(OAc)4	4
RuBr(CO)3(eta-C3H5)	4
S-pyrrolidine-2-carbaldehyde	3
SL-J009-1	2
SPhosAuNTf2	0
Schwartz's reagent	2
Selectfluor	1
TEA	4
Tetrakis(dimethylamino)ethylen	2
Tosyl isocyanate	0
Tri(p-tolyl)phosphine	0
Trifluoromethanesulfonamide	1
Triisopropyl borate	0
Trimethyl borate	0
Trimethyl orthoacetate	4
Trimethylacetic acid	4
Trimethylmethoxysilane	0
Triphenylphosphine oxide	0
Tris(3,6-dioxahexyl)amine	2
Tris(4-methoxyphenyl)phosphine oxide	2
TurboGrignard	2
VANOL-B3	3
W(CO)5	2
WA30 basic resin	1
Wilkinson's catalyst	0
XPhos	2
Yb(OTf)3 immobilized on sodium propylsulfonate and phenyl group co-functionalized magnetic core?mesoporous silica shell composite	4
Yb(hfc)3(+)	3
YerE from Yersinia pseudotuberculosis	2
Zn(2+)*CF3O3S(1-)*C6H18NSi2(1-)	4
[(1,3-bis(2,4,6-trimethylphenyl)imidazol-2-ylidene)3Zn3(H)4(THF)] [BPh4]2	0
[(1,3-bis(diphenylphosphino)propane)Pd(H2O)2](BF4)2	0
[(1,5-cyclooctadiene)(OH)iridium(I)]2	0

"column1"	"Min*(Score)"
[(C6H6)(PCy3)(CO)RuH]+*BF4	4
[(DPEphos)Rh(COD)]BF4	0
[(R)-(+)-1,1'-bi(2-naphthol)]Ti(Oi-Pr)2	2
[(eta5-C5Me5)RuCl(mu2-SMe)2Ru(eta5-C5Me5)Cl]	4
[1,1'-bis(diphenylphosphino)ferrocene]nickel(II) chloride	4
[1,3-bis(2,6-diisopropyl-phenyl)imidazol-2-ylidene] silver(I) chloro	0
[1,3-bis(2,6-diisopropylphenyl)imidazol-2-ylidene]gold bis(trifluoromethanesulfonyl)imidate	0
[2,2]bipyridinyl	4
[3,5-bis(trifluoromethyl)phenyl]-3-{(2S)-3,3-dimethyl-1-[(triphenylphosphoranylidene)amino]butan-2-yl}thiourea	3
[AuCl(IPr)]	0
[Cd2(tren)2(dl-alaninato)](ClO4)3*H2O	0
[Cp(P-iPr3)Ru(CH3CN)2](1+)*B(C6F5)4(1-)	4
[Cp*Co(C6H6)][B(C6F5)4]2	0
[Cp*Rh(CH3CN)3](BF4)2	0
[Cp*Rh(CH3CN)3][SbF6]2	0
[D3]phosphoric acid	0
[D]-sodium hydroxide	0
[Fe(5,10,15,20-tetraphenylporphyrin)]BF4	4
[Fe(5,10,15-triphenylcorrole)]BF4	4
[Fe{N(SiMe3)2}2]2	4
[In(S,S)-iPr-pybox](OTf)3	0
[Ir(1,5-cyclooctadiene)2]triflate	4
[Ir(2-(2,4-difluorophenyl)-4-(trifluoromethyl)pyridine)2(5,5'-bis(trifluoromethyl)-2,2'-bipyridine)]PF6	4
[Ir(COD)2]BF4	0
[IrH2(thf)2(PPh2Me)2]PF6	0
[MoO2Cl2(dmf)2]	1
[Ni(dimethylglyoxime)Cl2]	4
[Rh(OH)(cod)]2	0
[Rh(dppe)]ClO4	0
[Rh(dppp)]BF4	0
[Rh(nbd)(R,R)-Me-Duphos]ClO4	0
[Rh2(S-BPTPI)4]*3H2O	0
[Ru(kappa1-OAc)(kappa2-OAc)(kappa3-1,1,1-tris(diphenylphosphinomethyl)ethane)]	3
[bis(acetoxy)iodo]benzene	3
[bis(trifluoromethanesulfonyl)imidate](triphenylphosphine)gold(I)	4
[iridium(CH2CHCH2)(C6H2(Cl)(NO2)COO)((R)-2,2'-bis(diphenylphosphino)-5,5'-dichloro-6,6'-dimethoxy-1,1-biphenyl)]	3
[ruthenium(II)(eta6-1-methyl-4-isopropyl-benzene)(chloride)(mu-chloride)]2	4
[{(R)-H8-BINOLate}Ti(O-i-Pr)2]x	2
acetaldehyde	4
acetamide	4

"column1"	"Min*(Score)"
acetic acid	4
acetic acid hydrazide	0
acetic anhydride	4
acetonitrile	4
acetophenone	4
acetyl chloride	4
acetylacetonatodicarbonylrhodium(I)	0
acetylhydroxamic acid	1
acidic ion exchange resin	0
acidic ion-exchange resin P-SO ₃ H	0
air	4
alkali hydroxide	0
allyl(cyclopentadiene)palladium(II)	0
alpha cyclodextrin	2
alpha-picoline borane	0
alpha-picoline-borane	0
alumina*KF supported on silica	1
alumina-supported iron(III) chloride	1
aluminium	0
aluminium oxide hydroxide	0
aluminium trichloride	0
aluminium tris(2,6-diphenylphenoxide)	0
aluminium(III) triflate	4
aluminum (III) chloride	0
aluminum oxide	0
aluminum tri-bromide	0
aluminum tri-tert-butoxide	0
amberlyst-15	0
aminosulfonic acid	0
ammonia	0
ammonium acetate	4
ammonium bicarbonate	0
ammonium bromide	4
ammonium cerium(IV) nitrate	2
ammonium chloride	4
ammonium fluoride	0
ammonium formate	0
ammonium hexafluorophosphate	1
ammonium hydroxide	0
ammonium iodide	4

"column1"	"Min*(Score)"
ammonium metavanadate	0
ammonium peroxydisulfate	4
aniline	4
antimony pentafluoride	2
aqueous extract of the tamarind fruits	2
askanite-bentonite clay	0
barium dihydroxide	0
barium hydroxide monohydrate	0
barium hydroxide octahydrate	0
barium manganate	0
barium permanganate	0
barium(II) hydroxide	0
barium(II) iodide	0
barium(II) oxide	0
bathophenanthroline	1
benzaldehyde	4
benzaldehyde dimethyl acetal	3
benzenesulfonic acid	0
benzo[1,3,2]dioxaborole	0
benzoic acid	4
benzotriazol-1-ol	4
benzotriazol-1-yloxy-tris-(pyrrolidino)-phosphonium hexafluorophosphate	3
benzotrifuroxan	1
benzylamine	3
benzylmagnesium chloride	1
benzyltriethylammonium	3
benzyltriethylammonium bromide	3
beta-Zeolite	0
biphenyl	2
bis(1,5-cyclooctadiene)diiridium(I) dichloride	0
bis(1,5-cyclooctadiene)iridium(I) tetrafluoroborate	1
bis(1,5-cyclooctadiene)nickel(0)	4
bis(1,5-cyclooctadiene)nickel(0)	4
bis(1,5-cyclooctadiene)rhodium(I) tetrafluoroborate	0
bis(1,5-cyclooctadiene)rhodium(I) trifluoromethanesulfonate	0
bis(1-methyl-1-phenylethyl)peroxide	0
bis(2,2'-diamino-1,1'-binaphthyl)-based chiral phosphoramidate	0
bis(2,2,6,6-tetramethylpiperidin-1-yl)magnesium-bis(lithium chloride) complex	4
bis(2,6-diisopropylphenyl)imidazol-2-ylidene	3
bis(acetonitrile)(1,5-cyclooctadiene)rhodium(I) tetrafluoroborate	0

"column1"	"Min*(Score)"
bis(acetylacetonate)nickel(II)	4
bis(benzonitrile)palladium(II) dichloride	0
bis(bis(trimethylsilyl)amido)zinc(II)	4
bis(cyclohexanyl)borane	2
bis(cyclopentadienyl)titanium dichloride	1
bis(dibenzylideneacetone)-palladium(0)	0
bis(dicyclohexylphosphino)methane	0
bis(diethylamino)cyclopropenium tetraphenylborate	0
bis(eta3-allyl-mu-chloropalladium(II))	0
bis(ethylene)rhodium(I) chloride dimer	0
bis(norbornadiene)rhodium(I) tetrafluoroborate	0
bis(pinacol)diborane	0
bis(tertbutylcarbonyloxy)iodobenzene	4
bis(tri-n-butyltin)	0
bis(tricarbonyl(eta-cyclopentadienyl)tungsten)	0
bis(trifluoromethane)sulfonimide lithium	0
bis(trifluoromethanesulfonyl)amide	2
bis(triphenylphosphine) palladium (II) acetate	0
bis(triphenylphosphine)copper(I) tetrahydroborate	0
bis(triphenylphosphine)nickel(II) chloride	4
bis-diphenylphosphinomethane	3
bis-triphenylphosphine-palladium(II) chloride	0
bis[-(R)-MeCH-O-CH2-2-yl-pyridine-6-yl-CH2-O-(R)-MeCH-]	4
bis[2-(diphenylphosphino)phenyl] ether	3
bis[dichloro(pentamethylcyclopenta-dienyl)iridium(III)]	0
bisacetonitrile[norbornadiene]rhodium(I) hexafluoroantimonate	0
bismuth(III) bromide	3
bismuth(III) chloride	3
bismuth(III) iodide	3
bismuth(III) trifluoromethanesulfonate	3
bis{rhodium[3,3'-(1,3-phenylene)bis(2,2-dimethylpropanoic acid)]}	0
borane pyridine	4
borane pyridine complex	4
borane tert-butylamine	1
borane-THF	0
borane/tetrahydrofuran	4
boric acid	4
boric acid tributyl ester	4
boric anhydride	0
boron tribromide	1

"column1"	"Min*(Score)"
boron trichloride	1
boron trifluoride	1
boron trifluoride diethyl etherate	1
boron trioxide	1
brominated hydroxymethylbenzoic acid resin	4
bromine	2
bromopentacarbonylmanganese(I)	0
brucine N-oxide	2
buta-1,3-diene	2
butyl magnesium bromide	1
caesium carbonate	0
calcium carbonate	0
calcium carbonate pentahydrate	0
calcium chloride	4
calcium hydride	0
calcium oxide	2
calcium sulfate	0
calcium(II) bis-(trifluoromethanesulfonimide)	2
calcium(II) trifluoromethanesulfonate	2
camphor-10-sulfonic acid	4
carbocationic species *B(C6F5)4(-)	2
carbon dioxide	4
carbon monoxide	4
carbon tetrabromide	4
carbon-SO3H	4
carbonic acid dimethyl ester	2
carbonochloridic acid 1-chloro-ethyl ester	1
carbonyl bis(hydrido)tris(triphenyl-phosphine)ruthenium(II)	0
carboxypolystyrene	2
cellulose sulphuric acid	0
cerium(III) chloride	2
cerium(III) chloride heptahydrate	2
cesium acetate	4
cesium fluoride	4
cesium hydroxide	0
cesium pivalate	4
cetyltrimethylammonium bromide	3
cetyltrimethylammonium chloride	3
chiral amino alcohol ligand	2
chiral bis(1-naphthyl)methyl-amine-derived ligand	2

"column1"	"Min*(Score)"
chiral bis-pyridino-18-crown-6	1
chiral camphor-derived [2.2.1] bicyclic sulfide	1
chiral catalyst	2
chiral deriv. of [2-thiabicyclohept-3-yl]bicycloheptanone	2
chiral dipeptide N-acylethylenediamine-based ligand	0
chiral phosphoramidate catalyst	2
chiral thiazolyl-L-threonine-derived catalyst	2
chiral triazolium salt	2
chloranil	0
chloro(1,3-bis(2,6-di-i-propylphenyl)imidazol-2-ylidene)gold(I)	4
chloro(1,5-cyclooctadiene)rhodium(I) dimer	0
chloro(triphenylphosphine)gold(I)	4
chloro-trimethyl-silane	2
chloro[1,3-bis(2,6-di-i-propylphenyl)imidazol-2-ylidene]copper(I)	0
chloro[1,3-bis(2,6-diisopropylphenyl)imidazol-2-ylidene]gold(I)	4
chloro[tris(2,3,4,5,6-pentafluorophenyl)phosphine]gold(I)	4
chlorobis(ethylene)rhodium(I) dimer	0
chlorodicarbonyl(eta5-pentaphenylcyclopentadienyl)ruthenium(II)	4
chloroform	3
chlorosulfonic acid	0
chlorosulfonic acid supported piperidine-4-carboxylic acid functionalized Fe3O4 nanoparticles	0
cholin hydroxide	4
chromium chloride	0
chromium dichloride	0
chromium tricarbonyl	0
chromium(VI) oxide	0
cis-dichlorobis(triphenylphosphine)platinum(II)	0
citric acid	3
clay	0
cobalt(II) bromide	0
cobalt(II) bromide-[1,2-bis(diphenylphosphino)ethane]	3
cobalt(II) chloride	0
cobalt(II) phthalocyanine	0
cobalt(III) acetylacetonate	0
copper	0
copper (I) acetate	2
copper (I) trifluoromethane sulfonate benzene	0
copper (II) trifluoroacetate hydrate	4
copper acetylacetonate	4
copper diacetate	4

"column1"	"Min*(Score)"
copper dichloride	4
copper(I) 3-methylsalicylate	2
copper(I) bromide	2
copper(I) bromide dimethylsulfide complex	2
copper(I) oxide	1
copper(I) trifluoromethanesulfonate toluene complex	2
copper(I) trifluoromethanesulfonate * 1/2 toluene	2
copper(II) 2-ethylhexanoate	4
copper(II) acetate monohydrate	4
copper(II) acetylacetonate	4
copper(II) bis(trifluoromethanesulfonate)	4
copper(II) chloride dihydrate	4
copper(II) ferrite	4
copper(II) iodide	4
copper(II) nitrate hexahydrate	4
copper(II) oxide	4
copper(II) sulfate	4
copper(I) chloride	2
copper(I) cyanide	2
copper(I) iodide	2
copper(II) bromide	4
copper(II) sulfate pentahydrate	4
cucurbituril	2
cyanoborane	0
cyclo-octa-1,5-diene	2
cyclohexanone	2
cyclohexylamine	4
cyclohexyldiphenylphosphine	3
cyclopentyl methyl ether	2
cyclopentylmagnesium chloride	1
d-Ipc2BH	2
d8-isopropanol	2
dacarbazine	2
decacarbonylirhenium(0)	0
deuteriated sodium hydroxide	0
di(n-butyl)(iodo)tin hydride	0
di-isopropyl azodicarboxylate	2
di-mu-bromobis(tri-tert-butylphosphino)dipalladium(I)	0
di-mu-chlorobis(norbornadiene)dirhodium(I)	0
di-n-butylboryl trifluoromethanesulfonate	2

"column1"	"Min*(Score)"
di-n-butyliodotin hydride	0
di-n-butylzinc	1
di-tert-butyl dicarbonate	0
di-tert-butyl peroxide	1
di-tert-butyl(1,1'-biphenyl)-2-ylphosphinegold(I)bis(trifluoromethanesulfonimide)	4
di-tert-butyl(methyl)phosphonium tetrafluoroborate salt	1
di[(eta-1,2,5,6)-1,5-cyclooctadiene]rhodium hexafluoroantimonate	0
diallyl 1,4-dihydro-2,6-dimethylpyridine-3,5-dicarboxylate	4
diammonium sulfide	4
diazomethyl-trimethyl-silane	0
dibenzo-18-crown-6	1
dibenzoyl peroxide	0
dibenzylamine	3
diboron trioxide	1
dibromoborane	4
dibutyl tin diiodide	0
dibutylamine	4
dibutylbis(cyclopentadienyl)zirconium	0
dibutyl dimethoxytin	1
dibutyltin chloride	0
dibutyltin diacetate	0
dicarbonylacetylacetonato rhodium (I)	0
dichloro bis(acetonitrile) palladium(II)	0
dichloro(1,1'-bis(diphenylphosphanyl)ferrocene)palladium(II)*CH2Cl2	0
dichloro(pentamethylcyclopentadienyl)rhodium (III) dimer	0
dichlorogallane	2
dichloromethylsilane	4
dicyclohexyl-(2',6'-dimethoxybiphenyl-2-yl)-phosphane	2
dicyclohexyl-carbodiimide	4
dicyclohexylboron chloride	1
dicyclohexylphenylphosphine	0
dicyclopentylboron trifluoromethanesulfonate	1
diethyl 2,6-dimethyl-1,4-dihydropyridine-3,5-dicarboxylate	4
diethyl chlorophosphate	0
diethylaluminum iodide	0
diethylamine	2
diethylamino-sulfur trifluoride	0
diethylazodicarboxylate	0
diethylzinc	1
dihydrogen peroxide	0

"column1"	"Min*(Score)"
diiodomethane	2
diisobutylaluminium acetylacetonate	1
diisobutylaluminium hydride	0
diisopinocampheylborane	0
diisopinocampheylborane	0
diisopropoxy(eta2-propene) titanium(II) complex	1
diisopropyl zinc	1
diisopropyl-carbodiimide	4
diisopropylamine	3
dilithium (R)-3,3'-diphenylbinaphtholate	0
dilithium tetra(tert-butyl)zincate	2
dimesitylmagnesium	4
dimethyl sulfoxide	4
dimethyl zinc(II)	1
dimethylaluminum chloride	0
dimethylfumarate	4
dimethylphenyl(4,4,5,5-tetramethyl-1,3,2-dioxaborolan-2-yl)silane	2
dimethylsulfide	4
dimethylsulfide borane complex	0
dimethylsulfide gold(I) chloride	4
diphenyl hydrogen phosphate	0
diphenyl hydrogen phosphite	0
diphenyl((R)-1-((S)-1-phenylethyl)aziridin-2-yl)methanol	4
diphenyl((S)-1-((S)-1-phenylethyl)aziridin-2-yl)methanol	4
diphenyl(methyl)phosphine	0
diphenyl-((S)-1-((S)-1-phenylethyl)aziridin-2-yl)-methanol	4
diphenylborinic acid	4
diphenylboronchloride	1
diphenylsilane	0
dipotassium hydrogenphosphate	1
dipotassium peroxodisulfate	1
dirhodium tetraacetate	0
dirhodium(II) tetrakis(perfluorobutyrate)	0
dirhodium(II) tetrakis<(3S)-phthalimido-2-piperidinonate>	0
dmap	4
dodecacarbonyl-triangulo-triruthenium	4
dysprosium	0
dysprosium(III) trifluoromethanesulfonate	2
epi-cinchonidine	2
epiCDT	2

"column1"	"Min*(Score)"
erbium triisopropoxide	0
ethanol	4
ethanolamine	4
ethoxy(potassiosulfanyl)methanethione	1
ethyl 2-[2-(4-chlorophenyl)-2-oxoethyl]sulfonylacetate	4
ethyl acetate	4
ethyl bromide	3
ethyl iodide	2
ethylacrolein	2
ethylaluminum dichloride	0
ethylene dibromide	3
ethylene glycol	4
ethylenediamine	0
ethylenediamine diacetate	4
ethylenediamine diacetic acid	4
ethylenediaminediacetic acid	4
ethylenediaminetetraacetic acid	4
ethylmagnesium bromide	1
ethylmagnesium chloride	1
ethyltriphenylphosphonium bromide	1
europium(III) trifluoromethanesulfonate	2
ferric(III) bromide	4
fluoride	0
fluorous reverse-phase silica	1
formaldehyd	4
formamide	4
formic acid	4
furan	4
gadolinium(III) isopropoxide	0
gadolinium(III) trifluoromethanesulfonate	0
gallium(III) trichloride	4
gallium(III) triflate	4
germanium(II) chloride dioxane	4
girard's reagent T	0
glycine	4
gold bromide	4
gold(I) chloride	4
gold(III) bromide	0
gold(III) chloride	0
gold(III) acetate	0

"column1"	"Min*(Score)"
gold-on-silver film	4
graphene?mesoporous anatase nanocomposite	2
guanidine hydrochloride	4
hex-3-yne	2
hexafluorophosphoric acid	0
hexamethyldisilathiane	0
hydrazine	0
hydrazine hydrate	0
hydrazinium sulfate	0
hydrochloric acid diethyl ether	0
hydrogen	0
hydrogen bromide	0
hydrogen cation	0
hydrogen fluoride	0
hydrogen sulfide	0
hydroquinonein 1,4-phthalazinediyl diether	2
hydroquinone 2,5-diphenyl-4,6-pyrimidinediyl diether	2
hydroquinone	2
hydroxyapatite-encapsulated-gamma-Fe2O3 supported sulfonic acid nanoparticles	0
hydroxylamine	0
hydroxylamine acetate	4
hydroxylamine hydrochloride	0
hydroxylamine potassium salt	1
hypophosphorous acid	0
i-Pr2Et	2
iPr2NCOO (R)-2-[(S)-CH(4-Br-C6H4)OH]-2-Bu-cC6H8-(Z)=CH ester	2
iPr2NCOO (R)-2-[(S)-CH(C6H5)OH]-2-Bu-cC6H8-(Z)=CH ester	2
immobilized Co(II) Schiff base complex supported on multi?wall carbon nanotubes	0
indium	0
indium (III) iodide	0
indium iodide	0
indium tribromide	0
indium(I) bromide	0
indium(II) bromide	0
indium(III) bromide	0
indium(III) chloride	0
indium(III) triflate	4
iodine	2
iodosylbenzene	2
iron	0

"column1"	"Min*(Score)"
iron oxide	4
iron(II) chloride	4
iron(II) dodecylsulfate	4
iron(II) triflate	4
iron(II,III) oxide	4
iron(III) chloride	4
iron(III) chloride adsorbed on silica gel	4
iron(III) chloride hexahydrate	4
iron(III) paratoluenesulfonate	4
iron(III) trifluoromethanesulfonate	4
isopropyl alcohol	4
isopropyl bromide	2
isopropyl chloride	1
isopropyl magnesium chloride - lithium chloride complex	1
isopropyl magnesium lithium chloride	1
isopropyl lithium	0
isopropylmagnesium bromide	1
isopropylmagnesium chloride	1
lanthanum (III) chloride bis(lithium chloride) complex	4
lanthanum(III) triflate	4
lead	0
lead dioxide	0
lead(II) chloride	0
lead(II) iodide	0
lead(IV) acetate	0
lithium	0
lithium (10R)-9-dihydro-10-trimethylsilyl-9-borabicyclo[3.3.2]decane diethyl etherate	0
lithium (S)-B-H2-(10)-trimethylsilyl-9-borabicyclo[3.3.2]decane	0
lithium acetate	4
lithium aluminium tetrahydride	0
lithium borohydride	0
lithium bromide	4
lithium bromide monohydrate	4
lithium carbonate	0
lithium chloride	4
lithium cyanide	0
lithium di-n-butylcuprate	0
lithium dihydronaphthylide radical	0
lithium diisopropyl amide	0
lithium ethoxide	0

"column1"	"Min*(Score)"
lithium fluoride	4
lithium hexamethyldisilazane	0
lithium hydroxide	0
lithium hydroxide monohydrate	0
lithium iodide	4
lithium methanolate	0
lithium n-propoxide	0
lithium pentane-2,4-dionate	0
lithium tert-butoxide	0
lithium tetrafluoroborate	1
lithium triethylborohydride	0
lithium trifluoromethanesulfonate	0
lithium-B-H2-(10R)-trimethylsilyl-9-borabicyclo[3.3.2]decane	0
macroporous cyanoborohydride	0
magnesia	0
magnesium	0
magnesium bromide	4
magnesium bromide ethyl etherate	4
magnesium chloride	4
magnesium hydrogen sulfate	4
magnesium iodide	4
magnesium methanolate	0
magnesium oxide	4
magnesium sulfate	4
magnesium sulphate	4
magnesium triflate	4
maleic acid	0
manganese	0
manganese(II) chloride hexahydrate	0
manganese(II) sulfate	0
manganese(IV) oxide	0
manganese(II) chloride	0
mercaptoacetic acid	4
mercury dichloride	0
mercury(II) diacetate	0
mesitylcopper(I)	0
mesityllithium	0
meso-tetraphenylporphyrin iron(III) chloride	4
mesoporous aluminosilicate Al-MCM-41	1
mesoporous silica supported copper nano catalyst	1

"column1"	"Min*(Score)"
methanesulfonamide	1
methanesulfonic acid	0
methanesulfonyl chloride	0
methanol	4
methoxide	0
methoxy(cyclooctadiene)rhodium(I) dimer	0
methyl iodide	2
methyl magnesium iodide	1
methyl zinc (1+); methylate	0
methyl-((R)-1-phenyl-2-piperidin-1-yl-ethyl)-amine	2
methylamine	4
methylcyclopentadienyl manganese(I) tricarbonyl	0
methyllithium	0
methyllithium lithium bromide	0
methylmagnesium bromide	1
methylmagnesium chloride	1
methylthioninium chloride hydrate	0
methyltin(IV) trichloride	0
modified germanium	4
modified silica-supported 1-propyl-3-methyl- imidazolium?HSO4 catalyst	1
molecular sieve	0
montmorillonite K 10	1
montmorillonite K 10 clay	1
montmorillonite K-10	1
montmorillonite K10	1
montmorillonite K10 Clay	1
montmorillonite K10 clay	1
montmorillonite clay K10	1
morpholine	4
morpholinium acetate	4
mutant sperm whale myoglobin Mb(F43V,V68F)	2
n-Bu ₂ SnClH	0
n-butanethiol	2
n-butylammonium acetate	4
nano-Fe ₃ O ₄	4
naphthalen-1-yl-lithium	0
naphthalene	0
neodymium	0
neodymium(III) trifluoromethanesulfonate	1
neopentylmagnesium bromide	4

"column1"	"Min*(Score)"
nickel	0
nickel dibromide	4
nickel dichloride	4
nickel(II) acetate tetrahydrate	4
nickel(II) bromide dimethoxyethane	4
nickel(II) chloride hexahydrate	4
nickel(II) nitrate hexahydrate	4
nickel(II) perchlorate hexahydrate	4
nickel(II) triflate	4
nido-decaborane	0
niobium pentachloride	4
niobium(V) oxide	4
nitrobenzene	2
nitromethane	2
octylmagnesium bromide	1
ortho-(diphenyl-phosphino)-benzene sulphonic acid	0
ortho-diphenylphosphinobenzoic acid	3
ortho-ethylaniline	0
ortho-methylbenzoic acid	4
ortho-methylphenyl iodide	2
ortho-nitrobenzoic acid	4
orthoformic acid triethyl ester	0
oxalic acid	0
oxone	0
oxygen	2
oxygen-18	2
ozone	0
p-benzoquinone	2
palladium 10\% on activated carbon	1
palladium dichloride	0
palladium on activated carbon	1
palladium on activated charcoal	1
palladium on calcium fluoride poisoned with lead	0
palladium(II) chloride benzonitrile complex	0
palladium(II) hexafluoroacetylacetonate	0
palladium(II) iodide	3
palladium(II) trifluoroacetate	0
palladium(II)[(1,3-bis(diphenylphosphino)propane) (C6H5CN)2]*2BF4	0
pepsin from porcine gastric mucosa [EC 3.4.23.1]	2
per-rhenic acid	0

"column1"	"Min*(Score)"
perchloric acid	0
phenol	2
phenylborondichloride	1
phenylboronic acid	4
phenyllithium	0
phenylmagnesium bromide	1
phenylmagnesium chloride	1
phenylphosphinic acid	0
phenylsilane	0
phenyltrimethylammonium tribromide	3
phosphate buffer	4
phosphazene base-P4-tert-butyl	0
phosphomolybdic acid	1
phosphonic acid diethyl ester	0
phosphoric acid	0
phosphorus pentachloride	0
phosphorus pentoxide	0
phosphorus tribromide	0
phosphotungstic acid	0
phthalic anhydride	0
pi-allyl-palladium chloride	0
picoline-borane complex	2
pipecolic Acid	4
piperazine	2
piperdinium acetate	4
piperidin-2-one	4
piperidine	4
pivalaldehyde	4
platinum(II) chloride	0
platinum(IV) oxide	0
poly(methylhydrosiloxane)	2
polyethylene supported arsine	0
polymer-bound dimethylaminopyridine	1
polymer-bound trimethyl ammonium cyanoborohydride	1
polymer-supported 1,8-diazabicyclo[5.4.0]undec-7-ene	1
polymer-supported BH(OAc) ₃	1
polymer-supported chiral lithium amide	0
polymer-supported cyanoborohydride	1
polymethylhydrosiloxane	2
polyphosphoric acid	0

"column1"	"Min*(Score)"
polyphosphoric acid containing 84percent of P2O5	0
polystyrene cyanoborohydride	0
polystyrene-bound 4-(N-benzyl-N-methylamino)pyridine	1
polystyrene-bound super Broensted acid	2
polystyrene-supported sulfonic acid	0
potassium	0
potassium 2-methylbutan-2-olate	0
potassium 3,7-dimethyloctan-3-olate	0
potassium acetate	4
potassium bromide	4
potassium carbonate	0
potassium chloride	4
potassium cyanide	0
potassium diazodicarboxylate	2
potassium dihydrogenphosphate	1
potassium ethoxide	0
potassium fluoride	4
potassium fluoride 18-crown-6	1
potassium fluoride on aluminum oxide	1
potassium fluoride on basic alumina	1
potassium formate	1
potassium hexacyanoferrate(III)	4
potassium hexafluorophosphate	4
potassium hexamethylsilazane	3
potassium hydride	0
potassium hydrogen bifluoride	1
potassium hydrogen difluoride	1
potassium hydrogencarbonate	0
potassium hydrogenfluoride	1
potassium hydrogensulfate	1
potassium hydroxide	0
potassium iodide	4
potassium methanolate	0
potassium peroxymonosulfate	1
potassium phosphate	4
potassium sulfate	4
potassium tert-butylate	0
potassium tetrachloroaurate(III)	0
potassium thioacetate	4
potassium titanium oxalate dehydrate	2

"column1"	"Min*(Score)"
potassium triethylborohydride	0
potassium trifluoroacetate	4
potassium trimethylsilonate	1
praseodymium(III) isopropoxide	0
praseodymium(III) trifluoromethanesulfonate	2
propan-1-ol	2
propionic acid	4
propylamine	4
propylene diammonium diacetate	4
pyridine	4
pyridine N-oxide	4
pyridine hydrochloride salt	4
pyridine hydrogenfluoride	4
pyridinium chlorochromate	0
pyridinium p-toluenesulfonate	4
pyridinium triflate	4
pyrrole	4
pyrrolidine	3
quinindine	4
quinine	4
rac-Ala-OH	4
rac-Pro-OH	4
racemic BINOL derived phosphoric acid catalyst	0
racemic TBAT	0
resin Amberlyst A-31	1
rhenium(I) pentacarbonyl chloride	0
rhodium (II) octanoate dimer	0
rhodium(II) acetate	0
rhodium(II) pivalate	0
rhodium(III) chloride hydrate	0
rubidium hydroxide	0
ruphos	2
salicylic acid	4
samarium	0
samarium diiodide	2
scandium(III) acetate	2
sec.-butyllithium	0
selenium	0
silica gel	1
silicon carbide	4

"column1"	"Min*(Score)"
silicon-supported cyanoborohydride reagent	0
silicotungstic acid hydrate	0
silver	0
silver (II) carbonate	0
silver carbonate	0
silver fluoride	4
silver hexafluoroantimonate	2
silver nitrate	4
silver tetrafluoroborate	1
silver trifluoroacetate	4
silver trifluoromethanesulfonate	4
silver(I) acetate	4
silver(I) hexafluorophosphate	4
silver(I) triflimide	4
silver(I) oxide	4
silver-graphite	4
sodium	0
sodium (triacetoxy)borohydride	0
sodium acetate	4
sodium amalgam	0
sodium amide	0
sodium azide	3
sodium bis(2-methoxyethoxy)aluminium dihydride	0
sodium bis(trifluoromethanesulfonyl)imide	2
sodium borohydride acetate	0
sodium butanolate	0
sodium carbonate	1
sodium chloride	4
sodium chlorite	2
sodium cyanide	2
sodium cyanoborohydride	0
sodium cyanoborohydride resin	4
sodium cyanotrihydroborate	0
sodium deuterium cyanoborohydride	0
sodium diacetoxyl(acetyl)boranuide	0
sodium dihydrogen phosphate	4
sodium dihydrogen phosphate monohydrate	4
sodium dihydrogenphosphate	4
sodium disulfate	0
sodium disulfite	0

"column1"	"Min*(Score)"
sodium dithionate	4
sodium dithionite	0
sodium dodecyl-sulfate	4
sodium ethanolate	0
sodium formate	0
sodium hexafluoroantimonate	2
sodium hexamethyldisilazane	0
sodium hydride	0
sodium hydrogen sulfate	0
sodium hydrogencarbonate	0
sodium hydrogensulfite	0
sodium hydroxide	0
sodium iodide	0
sodium metabisulfite	0
sodium methoxide	0
sodium methylate	0
sodium nitrite	0
sodium ortho-iodobenzoate	2
sodium perborate	0
sodium perborate tetrahydrate	0
sodium periodate	0
sodium persulfate	0
sodium phenoxide	0
sodium phosphate	4
sodium pyrosulfate	4
sodium salt of sulphur oxide	2
sodium sulfate	4
sodium sulfide	0
sodium sulfite	0
sodium t-butanolate	0
sodium tert-pentoxide	0
sodium tetrachloroaurate(III) dihydrate	0
sodium tetrahydroborate	0
sodium tetrakis[(3,5-di-trifluoromethyl)phenyl]borate	0
sodium tri(benzoyloxy)borohydride	0
sodium triacetoxo borohydride	0
sodium triacetoxo borane hydride	0
sodium triacetoxo borohydride	0
sodium tris(acetoxy)borohydride	0
solid phase supported Sc(III)	2

"column1"	"Min*(Score)"
stannic bromide	0
succinic acid	2
sulfonated graphene	2
sulfonic acid supported on hydroxyapatite-encapsulated-gamma-Fe2O3	0
sulfur	0
sulfuric acid	0
suspension of male Wistar rat liver mitochondria	2
t r i s (4 , 4 ?-methoxydibenzylideneacetone)dipalladium(0)	0
t-butoxide	0
t-butyl dimethylsilyl triflate	4
tBu4ZnLi2	2
tbepc	2
tellurium	0
tert-Butyl peroxybenzoate	0
tert-butyl (2S,3R)-2-amino-3-hydroxybutanoate	2
tert-butyl alcohol	4
tert-butyl carbazate	2
tert-butylammonium hexafluorophosphate(V)	1
tert-butyl dimethylsilyl chloride	0
tert-butyl dimethylsilyl triflate	4
tert-butyl diphenylphosphine	3
tert-butyl diphenylsilyloxy 4-hydroxyproline	4
tert-butyl hypochlorite	0
tert-butyl isonitrile	0
tert-butyl magnesium chloride	1
tert.-butyl lithium	0
tert.-butyl hydroperoxide	0
tetra(4-chlorophenyl)porphyrin iron chloride	4
tetra(n-butyl)ammonium hydrogensulfate	2
tetra(n-butyl)ammonium hydroxide	0
tetra-(n-butyl)ammonium iodide	2
tetra-N-butylammonium tribromide	2
tetra-n-butylammonium fluoride trihydrate	0
tetrabutoxytitanium	2
tetrabutyl ammonium fluoride	0
tetrabutyl-ammonium chloride	1
tetrabutyl ammonium bromide	2
tetrabutyl ammonium acetate	4
tetrabutyl ammonium borohydride	0
tetrabutyl ammonium triphenyldifluorosilicate	4

"column1"	"Min*(Score)"
tetrabutylammonium triphenyldifluorostannate	0
tetrachlorobis(tetrahydrofuran)titanium(IV)	3
tetrachloromethane	4
tetrachlorosilane	2
tetraethoxy orthosilicate	4
tetraethylammonium fluoride	0
tetraethylammonium iodide	2
tetrafluoroboric acid	1
tetrafluoroboric acid diethyl ether	1
tetrafluoroboric acid diethyl ether complex	1
tetrahydrofuran	4
tetrakis(4-phenyl)methane-benzimidazole containing porous organic polymer	4
tetrakis(acetato)dimolybdenum(II)	1
tetrakis(acetonitrile)copper(I)tetrafluoroborate	1
tetrakis(acetonitrile)palladium(II) bis(tetrafluoroborate)	0
tetrakis(acetonitrile)palladium(II) tetrafluoroborate	0
tetrakis(actonitrile)copper(I) hexafluorophosphate	0
tetrakis(trifluoroacetato)rhodium(II)	0
tetrakis(triphenylphosphine) palladium(0)	0
tetramethylammonium chloride	4
tetramethoxymethane	4
tetramethylammonium triacetoxyborohydride	0
tetramethylenebis(magnesium chloride)	4
tetramethylpiperidyl MgCl LiCl	0
theophylline	4
thiamine diphosphate	4
thiazolium bromide	4
thionyl chloride	2
thiophene	4
thiophenol	2
tin	0
tin(II) chloride dihydrate	0
tin(II) iodide	0
tin(II) trifluoromethanesulfonate	0
tin(IV) chloride	0
tin(II) chloride	0
titanium tetra-n-propoxide	0
titanium tetrachloride	1
titanium tetrakispropoxide	0
titanium tris(diethylamido)chloride	2

"column1"	"Min*(Score)"
titanium(III) triisopropoxide	0
titanium(IV) bromide	2
titanium(IV) dichlorodiisopropylate	2
titanium(IV) iodide	2
titanium(IV) isopropylate	3
titanium(IV) tetraethanolate	0
titanium(IV) trichloride isopropoxide	0
titanium(IV)isopropoxide	0
titanocene(III) chloride	2
tol uene-4-sulfonic acid	0
toluene-4-sulfonic acid	0
toluene-4-sulfonic acid hydrazide	0
tri tert-butylphosphoniumtetrafluoroborate	1
tri-1-naphthylphosphine	0
tri-n-butyl-tin hydride	0
tri-n-butyltin lithium	0
tri-tert-butyl phosphine	0
triacetoxysodium borohydride	0
triacetoxylborane	2
tributyl borane	0
tributyl-amine	4
tributylphosphine	0
tricarbonylcyclopentadienyltungsten(II) chloride	0
trichloroacetic acid	4
trichloroacetonitrile	4
trichlorophosphate	0
trichlorosilane	0
tricyclohexylphosphine	0
tricyclohexylphosphine[1,3-bis(2,4,6-trimethylphenyl)-4,5-dihydroimidazol-2-ylidene][benzylidene]ruthenium(II) dichloride	4
tricyclopentylphosphine	0
triethanolamine	0
triethyl borane	0
triethyl borate	0
triethyl gallium	4
triethyl phosphate	0
triethyl phosphite	0
triethyl-sulfopropylammonium dihydrogen phosphomolybdate	1
triethylamine	4
triethylamine hydrochloride	4
triethylammonium methylpolystyrene triacetoborohydride	2

"column1"	"Min*(Score)"
triethylbutylammonium chloride	4
triethylphosphine	0
triethylsilane	3
triethylsilyl trifluoromethyl sulfonate	2
trifluoroacetic acid-d1	4
trifluoroacetic anhydride	4
trifluoroborane diethyl ether	2
trifluoromethylsulfonic anhydride	0
trifluoromethanesulfonic acid	0
trifuran-2-yl-phosphane	2
triisopropoxytitanium(IV) chloride	2
triisopropyl phosphite	0
triisopropylsilyl chloride	0
trimethoxysilane	0
trimethyl orthoformate	4
trimethylaluminum	0
trimethylamine	4
trimethylamine-N-oxide	0
trimethylphenylsilane	0
trimethylphosphane	3
trimethylsilyl acetate	4
trimethylsilyl bromide	0
trimethylsilyl iodide	0
trimethylsilyl trifluoromethanesulfonate	0
trimethylsilylazide	0
trimethylsilylmethyl lithium	0
trimethylsilylphosphate	1
trimethylsilyltributyltin	0
triphenyl phosphite	1
triphenyl-arsane	2
triphenylacetic acid	4
triphenylborane	0
triphenylphosphine on polystyrene	0
triphenylsilyl perrhenate	0
tris hydrochloride	4
tris(2,4-di-tert-butylphenyl)phosphite gold(I) chloride	4
tris(acetonitrile)(eta5-pentamethylcyclo-pentadienyl)rhodium(III) hexafluoroantimonate	0
tris(dibenzylideneacetone)dipalladium (0)	0
tris(dibenzylideneacetone)dipalladium(0) chloroform complex	0
tris(dimethylamino) sulphonium bifluoride	0

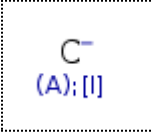


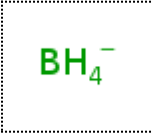


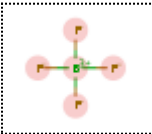
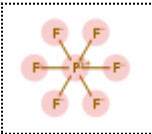
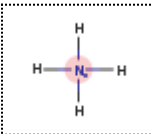
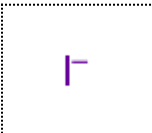
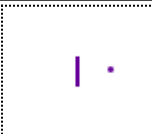
"column1"	"Min*(Score)"
tris(dimethylamino)sulfonium trimethylsilyldifluoride	0
tris(ethoxy)monochloro titanium	2
tris(methoxyethoxyethyl)amine	4
tris(p-bromophenylammoniumyl) hexachloroantimonate	2
tris(pentafluorophenyl)borate	1
tris-(2,2'-bipyridine)ruthenium(II) chloride	4
tris-(2-carboxyethyl)-phosphine hydrochloride	4
tris-(dibenzylideneacetone)dipalladium(0)	0
tris-(m-sulfonatophenyl)phosphine	0
tris-(o-tolyl)phosphine	0
tris-(triphenylsiloxy)-vanadium oxide	0
trityl tetrakis(pentafluorophenyl)borate	0
tungstosilicic acid hydrate	0
urea	3
vanadyl acetylacetonate	0
vanadyl triflate	4
water	4
water-d2	4
ytterbium(III) triflate	4
ytterbium(III) trifluoromethanesulfonate hydrate	3
ytterbium(III) trifluoromethanesulfonate nonohydrate	3
yttrium(III) chloride	0
yttrium(III) trifluoromethanesulfonate	0
yttrium(III) nitrate hexahydrate	0
zinc	0
zinc acetate dehydrate	4
zinc chloride diethyl ether	2
zinc diacetate	4
zinc dibromide	4
zinc dichloro(N,N,N',N'-tetramethylethylenediamine)	4
zinc trifluoromethanesulfonate	4
zinc(II) chloride	4
zinc(II) hydroxide	0
zinc(II) iodide	4
zinc(II) oxide	4
zinc(II) sulfate	4
zinc(II) tetrahydroborate	4
zirconium (IV) butoxide	0
zirconium complex of (R)-3,3'-diiodo-BINOL on 3 A MS	0
zirconium triflate	4

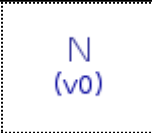
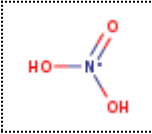
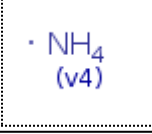
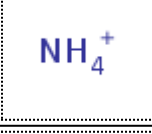
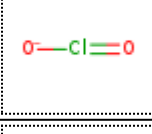
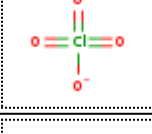
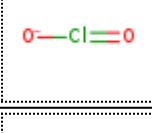
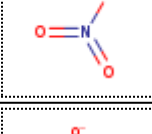

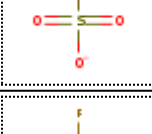
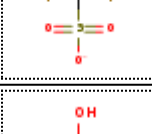
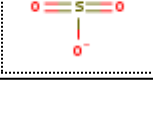
"column1"	"Min*(Score)"
zirconium(IV) chloride	0
zirconium(IV) tert-butoxide	0
zirconocene dichloride	0
{(2-methyl-2-phenyl-propylidene)((2,6-dimethylphenyl)imido)molybdenum(VI)bis(hexafluoro-tert-butoxide)}	0
{(2-methyl-2-phenyl-propylidene)((2,6-dimethylphenyl)imido)molybdenum(VI)bis(hexafluoro-tert-butoxide)}	0
{(R)-H8-BINOL}Ti(O-i-Pr) ₂	3
(S)-3,3'-bis(4''-trifluoromethylphenyl)-2,2'-(2,2-bisbromo-2-stannapropane-1,3-diyl)-1,1'-binaphthyl	1
(1,4,7,10-tetraoxacyclododecane)	4
(1,3,5-triaza-7-phosphaadamantane)	3
(3R,5R,7R)-N-((S)-(6-methoxyquinolin-4-yl)((1S,2S,4S,5R)-5-vinylquinuclidin-2-yl)methyl)adamantane-1-sulfamide	3
(R)-(+)-(4,4'-bi-1,3-benzodioxole)-5,5'-diylbis(di(3,5-dimethylphenyl)phosphine)	2
(R)-2-(piperidin-1-yl)-1,1,2-triphenylethanol	4
1,2-bis-(dicyclohexylphosphino)ethane	4
1,3,5-trichloro-2,4,6-triazine	4
1,3-bis-(2,6-diisopropylphenyl)-imidazol-2-ylidene	4
1-(2,6-diisopropylphenyl)-3-(3-sulfonatopropyl)imidazolium	3
1-ethyl-(3-(3-dimethylamino)propyl)-carbodiimide hydrochloride	4
1-ethyl-3-methylimidazolium hexafluorophosphate	3
2,2'-azobis(isobutyronitrile)	2
2,2'-iminobis[ethanol]	4
2,2-bis[(4S)-4-isopropylloxazolin-2-yl]propane	3
2,3-diazobicyclo[2.2.1]heptane bis-hydrochloride	4
2,3-dimethyl-but-1,3-diene	3
2,4,6-triisopropyl-N-((S)-(6-methoxy-2-phenylquinolin-4-yl)((1S,2S,4S,5R)-5-vinylquinuclidin-2-yl)methyl)benzenesulfonamide	2
2,6-di-tert-butyl-4-methyl-phenol	4
3-butyl-1,2-dimethyl-1H-imidazol-3-ium hydroxide	3
3-methyl-N-(3-methylbutyl)-1-butanamine	4
3-quinuclidinol	4
4-(benzyloxy)-N-(2-{1-[(4-fluorophenyl)methyl]piperidin-4-yl}ethyl)benzamide	3
4-chloro-benzenesulfonic acid	2
4-chlorobenzophenone	4
4-hydroxy-4-[2-(2,2,2-trifluoro-ethanoyl-amino)-ethyl]-piperidine-1-carboxylic acid tert-butyl ester	4
9-bora-bicyclo[3.3.1]nonane	3
9-borabicyclo[3.3.1]nonane	3
9-borabicyclo[3.3.1]nonane dimer	3
Acetyl bromide	4
(1R,2R)-1,2-bis[(2-diphenylphosphanyl)benzoylamino]cyclohexane	2

"column1"	"Min*(Score)"
1,4-dichlorocyclohexane	3
p-toluenesulfonic acid monohydrate	2
p-toluenesulfonyl chloride	3
triphenylphosphine	2

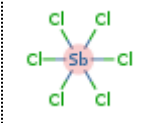

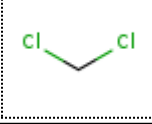

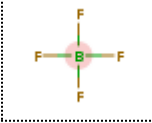
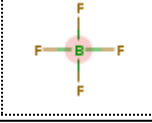

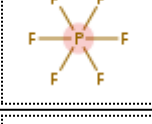
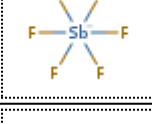
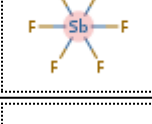
KNIME Report II - salt table

Knime report powered by Birt

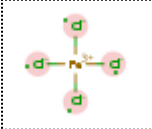
"salt string"	"salt smiles"
[Cl-]	
[Br-]	
[Br]	
[BH4-]	
[Cl]	
[F-]	
[F-][B+3]([F-])([F-])[F-]	
[F-][P+5]([F-])([F-])([F-])([F-])[F-]	
[H][N]([H])([H])[H]	
[I-]	
[I]	

"salt string"	"salt smiles"
[N]	
[N](=O)(O)O	
[NH4]	
[NH4+]	
[O-][Cl]=O	
[O-]Cl(=O)(=O)=O	
[O-]Cl=O	
[O-]N(=O)=O	
[O-]P([O-])([O-])=O	
[O-]S([O-])(=O)=O	
[O-]S(=O)(=O)C(F)(F)F	
[O-]S(O)(=O)=O	

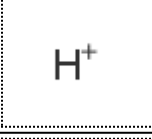
"salt string"	"salt smiles"
[O-2]	
[OH-]	
[O]	
[P](=O)(O)(O)O	
[P](F)(F)(F)(F)(F)F	
[S-2]	
[S](=O)(=O)(O)O	
B	
Br	
C[Si](C)(C)[O-]	
Cl	
Cl.Cl	

"salt string"	"salt smiles"
<chem>Cl[Sb](Cl)(Cl)(Cl)Cl</chem>	
<chem>Cl[Sn](Cl)(Cl)Cl</chem>	
<chem>ClCCl</chem>	
<chem>F</chem>	<chem>HF</chem>
<chem>F[B-](F)(F)F</chem>	
<chem>F[B](F)(F)F</chem>	
<chem>F[B](F)(F)F</chem>	
<chem>F[P-](F)(F)(F)F</chem>	
<chem>F[P](F)(F)(F)F</chem>	
<chem>F[Sb-](F)(F)(F)F</chem>	
<chem>F[Sb](F)(F)(F)F</chem>	
<chem>I</chem>	<chem>HI</chem>

"salt string"	"salt smiles"
N	
NO	
O	
O[Cl](=O)(=O)=O	
O=S(=O)([N-]S(=O)(=O)C(F)(F)F)C(F)(F)F	
O=S(=O)([N-]S(=O)(=O)C(F)(F)F)C(F)(F)F	
O=S(=O)(O)O	
OC(=O)C(F)(F)F	
OCl(=O)(=O)=O	
ON(=O)=O	
S	
[Ag+]	

"salt string"	"salt smiles"
[57Fe++]	${}^{57}\text{Fe}^{2+}$
[Al+3]	Al^{3+}
[Ba+2]	Ba^{2+}
[Bi+3]	Bi^{3+}
[Ca]	Ca
[Ca++]	Ca^{2+}
[Cd]	Cd
[Cd+]	Cd^+
[Cd+2]	Cd^{2+}
[Ce]	Ce
[Ce+4]	Ce^{4+}
[Cl-][Fe+3]([Cl-])([Cl-])[Cl-]	

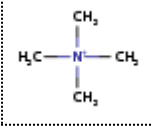
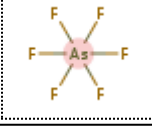
"salt string"	"salt smiles"
[Co]	Co
[Co++]	Co ²⁺
[Co+2]	Co ²⁺
[Co+3]	Co ³⁺
[Cs+]	Cs ⁺
[Cu]	Cu
[Cu+]	Cu ⁺
[Cu++]	Cu ²⁺
[Cu+2]	Cu ²⁺
[Er+3]	Er ³⁺
[Eu+3]	Eu ³⁺
[Fe]	Fe


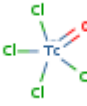


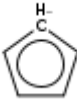
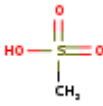
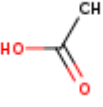

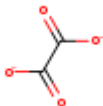
"salt string"	"salt smiles"
[Fe++]	
[Fe+2]	
[Fe+3]	
[Ga+2]	
[Ga+2]Ga+2]	
[Gd+3]	
[H]	
[H+]	
[H-]	
[I-][Zn++](I-)(I-)[I-]	
[I-][Zn+2]1(I-)[I-][Zn+2](I-)(I-)[I-]1	
[Ir+3]	

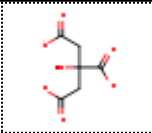
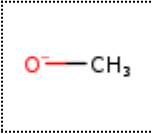

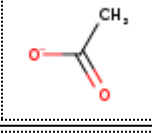
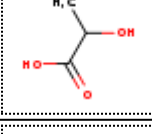
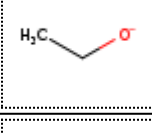
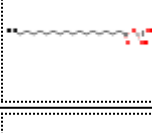
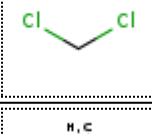
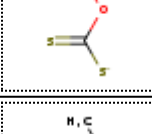
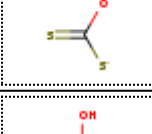
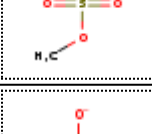
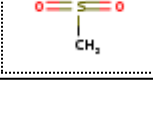
"salt string"	"salt smiles"
[K]	K
[K+]	K ⁺
[KH]	KH
[La]	La
[La+3]	La ³⁺
[Li]	Li
[Li+]	Li ⁺
[LiH]	LiH
[Lu+3]	Lu ³⁺
[Mg]	Mg
[Mg++]	Mg ²⁺
[Mg+2]	Mg ²⁺

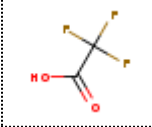
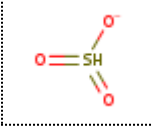
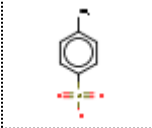
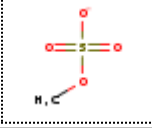
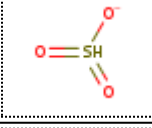
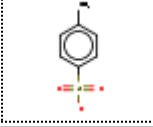
"salt string"	"salt smiles"
[Mn]	Mn
[Mn+2]	Mn ²⁺
[Mo]	Mo
[Mo+6]	Mo ⁶⁺
[Mo+6][Mo+6]	Mo ⁶⁺ —Mo ⁶⁺
[Na]	Na
[Na+]	Na ⁺
[NaH]	NaH
[Ni]	Ni
[Ni+2]	Ni ²⁺
[Pb+2]	Pb ²⁺
[Pd]	Pd

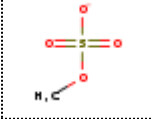
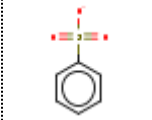
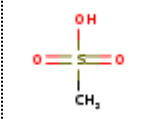
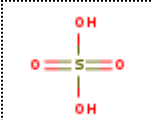
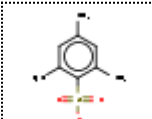


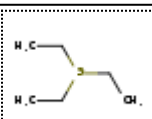
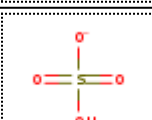
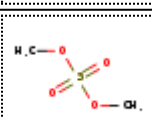
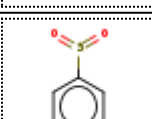
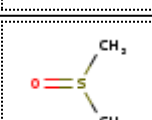
"salt string"	"salt smiles"
[Pd++]	Pd^{2+}
[Pd+2]	Pd^{2+}
[Pt+2]	Pt^{2+}
[Re+]	Re^{+}
[Ru]	Ru
[Sn]	Sn
[Sn+4]	Sn^{4+}
[Sr]	Sr
[SrH2]	SrH_2
[Ta+5]	Ta^{5+}
[Tc+2]	Tc^{2+}
[Ti]	Ti

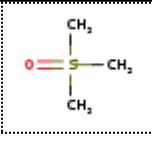
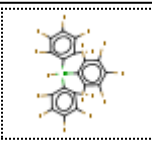
"salt string"	"salt smiles"
[Ti+4]	Ti^{4+}
[Tm+3]	Tm^{3+}
[Yb+3]	Yb^{3+}
[Zn]	Zn
[Zn]	Zn
[Zn+2]	Zn^{2+}
[Zr+4]	Zr^{4+}
C[N+](C)(C)C	
F[As](F)(F)(F)F	
Mo	?
Ni	?
O=[Mo++]=O	$O=Mo^{2+}=O$

"salt string"	"salt smiles"
<chem>O=[Mo+2]=O</chem>	
<chem>O=[Tc-](Cl)(Cl)(Cl)Cl</chem>	
<chem>O=[U+2]=O</chem>	
<chem>O=[V+2]</chem>	
<chem>[C]</chem>	<chem>C</chem> (v0)
<chem>[cH-]1cccc1</chem>	
<chem>[CH3][S](=O)(=O)(O)</chem>	
<chem>[CH3]C(=O)O</chem>	
<chem>[O-]C</chem>	<chem>H3C-O-</chem>
<chem>[O-]C=O</chem>	
<chem>C</chem>	<chem>CH4</chem>
<chem>C(=O)(C(=O)[O-])[O-]</chem>	

"salt string"	"salt smiles"
<chem>C(C(=O)[O-])C(CC(=O)[O-])(C(=O)[O-])O</chem>	
<chem>C[O-]</chem>	
<chem>c1cc([CH3])ccc1[S](=O)(=O)(O)</chem>	
<chem>CC(=O)[O-]</chem>	
<chem>CC(O)C(=O)O</chem>	
<chem>CC[O-]</chem>	
<chem>CCCCCCCCCCCCCCCCCC(O[Al](O)O)=O</chem>	
<chem>ClCCl</chem>	
<chem>COC([S-])=S</chem>	
<chem>COC([S-])=S</chem>	
<chem>COS(O)(=O)=O</chem>	
<chem>CS([O-])(=O)=O</chem>	

"salt string"	"salt smiles"
<chem>FC(F)(F)C(=O)O</chem>	
[Dy]	Dy
[Zn++]	Zn ²⁺
[Cl-]	Cl ⁻
[Hg++]	Hg ²⁺
[Hg]	Hg
[Dy+++]	Dy ³⁺
<chem>S([O-])(=O)=O</chem>	
<chem>Cc1ccc(cc1)S([O-])(=O)=O</chem>	
<chem>COS([O-])(=O)=O</chem>	
<chem>S([O-])(=O)=O</chem>	
<chem>Cc1ccc(cc1)S([O-])(=O)=O</chem>	

"salt string"	"salt smiles"
<chem>COS([O-])(=O)=O</chem>	
<chem>[O-]S(=O)(=O)c1ccccc1</chem>	
<chem>CS(O)(=O)=O</chem>	
<chem>OS(O)(=O)=O</chem>	
<chem>Cc1cc(C)c(c(C)c1)S([O-])(=O)=O</chem>	
<chem>OS(=O)(=O)C(F)(F)F</chem>	
<chem>Cc1ccc(cc1)S(O)(=O)=O</chem>	
<chem>CC[S+](CC)CC</chem>	
<chem>OS([O-])(=O)=O</chem>	
<chem>COS(=O)(=O)OC</chem>	
<chem>O=[S-](=O)c1ccccc1</chem>	
<chem>CS(C)=O</chem>	

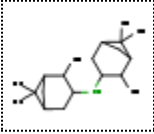
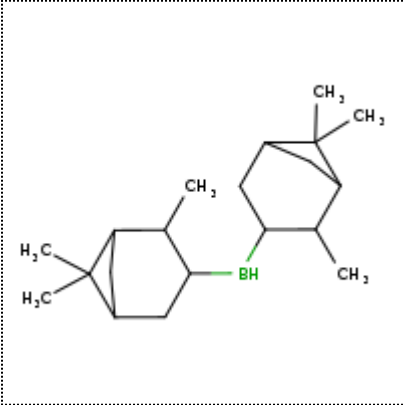
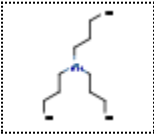
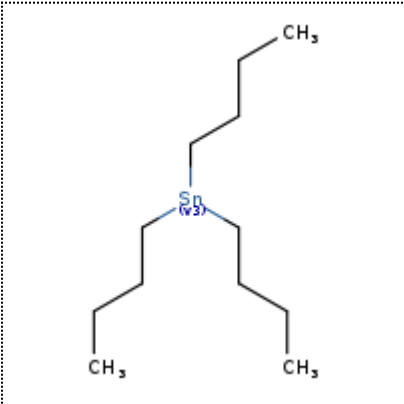
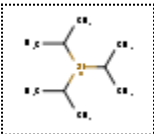
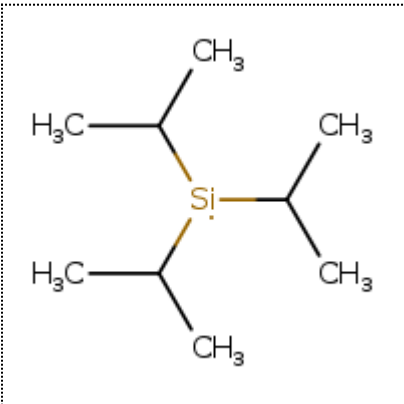
"salt string"	"salt smiles"
C[S+](C)(C)=O	
Fc1c(F)c(F)c(c(F)c1F)[B-](F) (c1c(F)c(F)c(F)c(F)c1F)c1c(F)c(F)c(F)c(F)c1F	

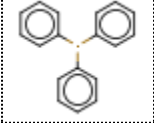
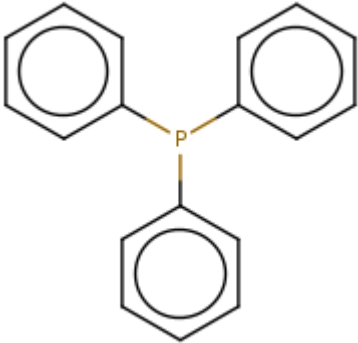

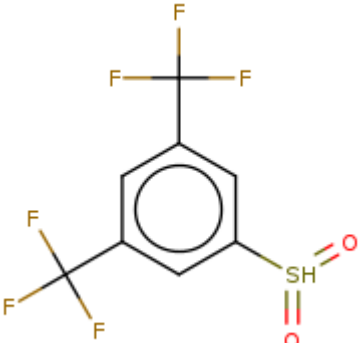

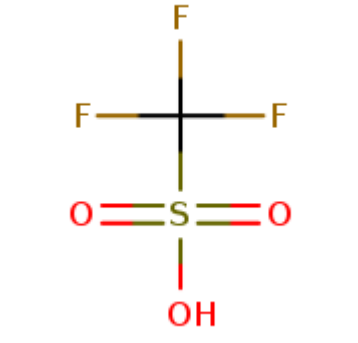
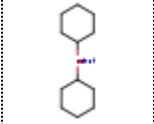
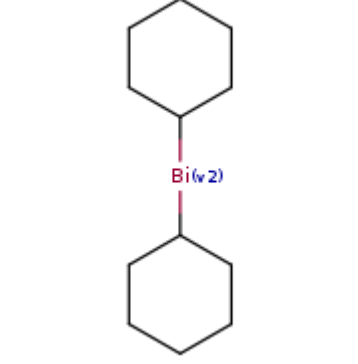
Reference reactions as landmarks

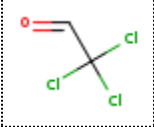
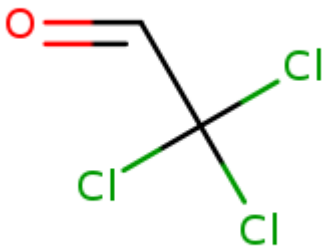
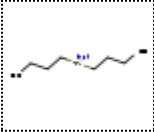

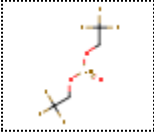
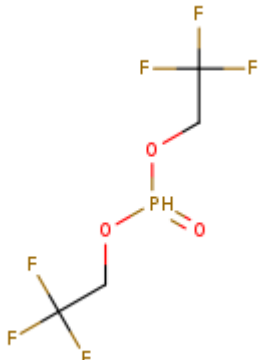
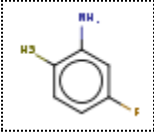
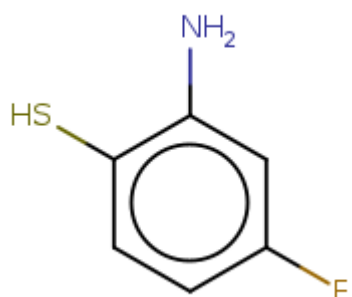
BIGINELLI	
POVAROV	
AZA-DIELS-ALDER	
POVAROV_2	
SNAP	
UGI	
GROEBKE-BLACK...	
GEVORGYAN	
PICTET-SPENGLER	
CUSHMAN	
PETASIS	

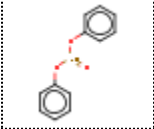
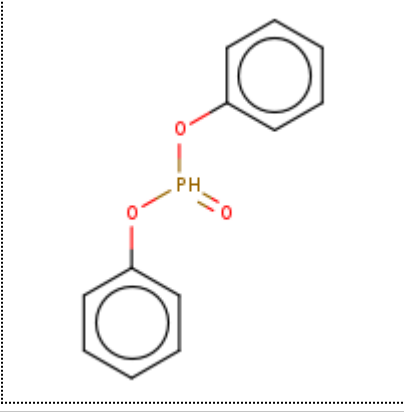
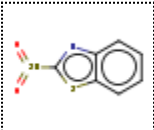
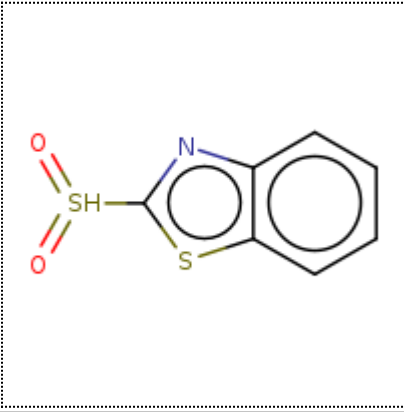
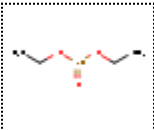
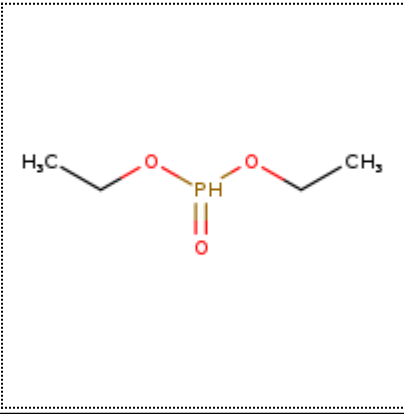

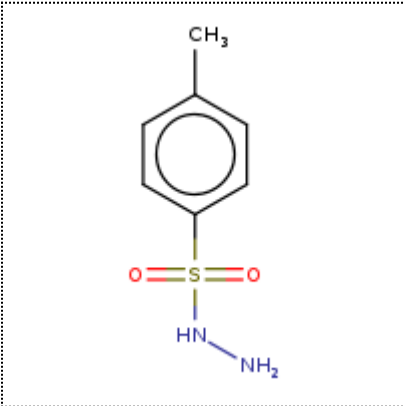
KNIME Report IV - LGs list

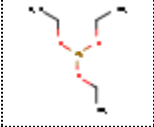
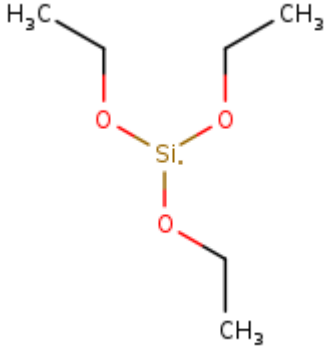
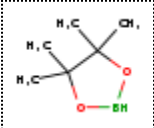
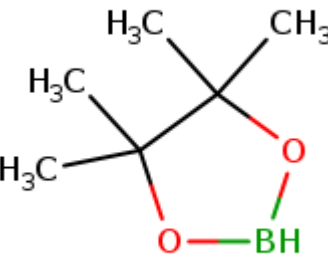
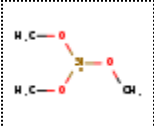
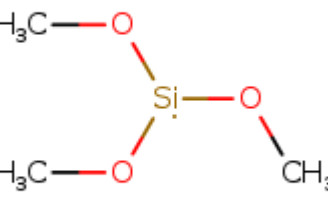
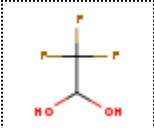
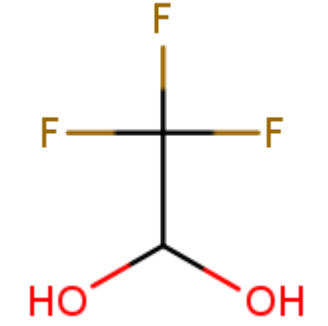
Knime report powered by Birt

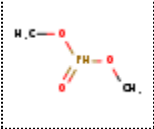
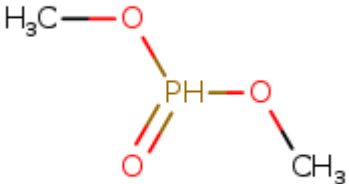
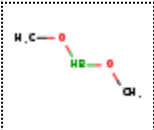
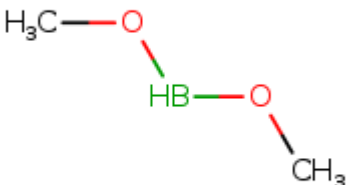
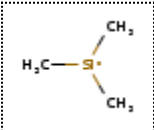
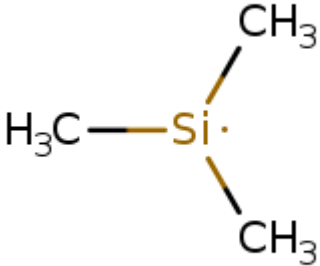
"LGs"	"LGs rendered with Marvin"
	
	
	

"LGs"	"LGs rendered with Marvin"	
		
		
		
		

"LGs"	"LGs rendered with Marvin"	
		
		
		
		

"LGs"	"LGs rendered with Marvin"
	
	
	
	

"LGs"	"LGs rendered with Marvin"	
		
		
		
		

"LGs"	"LGs rendered with Marvin"	
 <p>A small chemical structure of dimethyl phosphinic acid, showing a central phosphorus atom (P) double-bonded to an oxygen atom (O) and single-bonded to two methoxy groups (H₃C-O).</p>		 <p>A larger chemical structure of dimethyl phosphinic acid, showing a central phosphorus atom (P) double-bonded to an oxygen atom (O) and single-bonded to two methoxy groups (H₃C-O).</p>
 <p>A small chemical structure of dimethyl phosphinic acid, showing a central phosphorus atom (P) double-bonded to an oxygen atom (O) and single-bonded to two methoxy groups (H₃C-O).</p>		 <p>A larger chemical structure of dimethyl phosphinic acid, showing a central phosphorus atom (P) double-bonded to an oxygen atom (O) and single-bonded to two methoxy groups (H₃C-O).</p>
 <p>A small chemical structure of trimethylsilane, showing a central silicon atom (Si) single-bonded to three methyl groups (H₃C).</p>		 <p>A larger chemical structure of trimethylsilane, showing a central silicon atom (Si) single-bonded to three methyl groups (H₃C).</p>