UNIVERSITÀ DEGLI STUDI DI TRIESTE

XXX CICLO DEL DOTTORATO DI RICERCA IN NANOTECNOLOGIE

A multiscale methodology for the preliminary screening of alternative process designs from a sustainability viewpoint adopting molecular and process simulation along with data envelopment analysis

Settore scientifico-disciplinare: **ING-IND/24**

DOTTORANDO / A
ANDREA MIO

COORDINATORE
PROF. LUCIA PASQUATO

SUPERVISORE DI TESI
PROF. MAURIZIO FERMEGLIA

ANNO ACCADEMICO 2016/2017

Dedicated to

Anna

for the support she gave me

during this incredible journey

SUMMARY

Research activity in chemical engineering is focused on the refinement of theories and techniques employed for the development of new tools aiming at solving issues directly related to the generation of goods and services supplied by chemical, biochemical and pharmaceutical industries. Meanwhile, the rise of computational capabilities enabled scientists to include simulation techniques within the valuable methods to deal with a plethora of purposes. Among the others, multiscale approaches revealed to be very useful, since they broaden the *in-silico* application perspectives embracing theories from quantum mechanics at the nanoscale to classical mechanics at the macroscale.

Furthermore, the acknowledgment of sustainability among the cornerstones of future development led to a copious diffusion of sustainability evaluation methodologies. Henceforth, economic, social and environmental concerns have become pivotal within chemical processes assessments.

In this context, this thesis deals with the development of a multiscale framework for the preliminary screening of chemical process designs, promoting the adoption of various computational tools along with sustainability considerations.

The purpose of this methodology resides in the fulfillment of an emblematic need for any production site, i.e. the evaluation of a production process considering possible modifications from different perspectives in order to identify as fast as possible the most efficient design including economic, social and environmental concerns.

The reader will be guided through this topic following the chapters of the present dissertation.

In *Chapter I*, the concept of sustainability and sustainable development will be presented, along with several applications from various perspectives. Starting from the wider panorama of international institutions, the focus will sharpen towards industry concerns, concluding with some relevant examples from chemical process engineering field.

Chapter II will describe the steps to be performed to achieve the sustainability evaluation of the process alternatives. The identification of promising process designs is followed by the implementation of each flowsheet in a process simulator in order to calculate several indicators based on the sustainability pillars. The indicators scores will become the input of a mathematical tool (Data Envelopment Analyis) which aims to select the most efficient designs. The last step involves the employment of a retrofit analysis in order to identify the major sources of impacts, allowing the selection of the most suitable parameters to be tuned in the perspective of sub-optimal designs enhancement.

Chapter III will deal with the application of different molecular simulation techniques for the assessment of the octanol-water partition coefficient (Kow), which is a pivotal parameter for the calculation of several sustainability indicators.

Three case studies will be described in *Chapter IV*. The first one belongs to the pharmaceutical field and deals with the production of an antidiabetic drug (pioglitazone hydrochloride) considering different synthesis routes from various patents. The second case study regards the biochemical industry, dealing with the optimization of the operating conditions of a reactor employed for the production of biodiesel from vegetable oil. The last one explores the synthesis of a nanomaterial, adopting a sustainable approach for the evaluation of several reaction parameters involved in the production of a nanostructured semiconductor, i.e. cadmium selenide (CdSe) quantum dots.

Some concluding remarks and future perspectives will be included in the final *Chapter V*.

RIASSUNTO

La ricerca scientifica nell'ambito dell'ingegneria chimica si è focalizzata sia sul perfezionamento delle teorie e delle tecniche utilizzate attualmente, che sullo sviluppo di nuovi strumenti atti a risolvere le problematiche ancora insolute relative alle produzioni di beni e servizi tipici delle industrie chimiche, biochimiche e farmaceutiche. In contemporanea, il progressivo aumento delle potenzialità del calcolo computazionale ha permesso l'utilizzo di tecniche di simulazione *in-silico* in una grande varietà di applicazioni. Gli approcci multiscala si sono rivelati molto utili grazie alla loro peculiarità di coniugare aspetti che spaziano dalla quanto-meccanica tipica della nanoscala, alla meccanica classica dei materiali massivi, comprendendo prospettive molto ampie e adattando ogni teoria alle diverse applicazioni.

Inoltre, il riconoscimento dei concetti legati alla sostenibilità come principi cardine per ottenere uno sviluppo sostenibile ha generato un prolifico incremento della diffusione di metodologie per considerare aspetti sociali e ambientali, a fianco delle tradizionali stime economiche, nel quadro più ampio delle valutazioni degli impianti chimici.

In questo contesto, questa tesi tratta dello sviluppo di una metodologia multiscala per la stima preliminare di diverse configurazioni impiantistiche, promuovendo l'adozione di strumenti computazionali differenti e comprendendo valutazioni di carattere economico, sociale e ambientale.

Il fine ultimo che tale metodologia si prefigge risiede nella soddisfazione della necessità tipica di qualsiasi impianto di produzione, ovvero nella definizione di una metodologia di valutazione di vari parametri e configurazioni impiantistiche, utilizzando un'ottica sostenibile e fornendo risultati velocemente.

Al lettore verranno fornite le adeguate informazioni sull'argomento in maniera progressiva attraverso i capitoli di questa tesi.

Nel *Chapter I* saranno descritti il concetto di sostenibilità e di sviluppo sostenibile. Seguirà una trattazione riguardante la loro applicazione nella società odierna da diverse prospettive: a partire da quella più generalista delle istituzioni, fino a quella più particolare dell'industria, per concludere con una parte specifica sull'industria chimica, corredata di esempi di metodologie applicate a processi chimici.

Il *Chapter II* descriverà i passaggi necessari ad ottenere la valutazione della sostenibilità delle alternative impiantistiche. Dal reperimento delle informazioni necessarie, all'implementazione dei modelli nei simulatori di processo, seguito dal calcolo degli indici rappresentativi dei pilastri della sostenibilità, i cui valori vengono successivamente valutati tramite un algoritmo matematico (DEA) per identificare la configurazione impiantistica ottimale. Infine è necessario analizzare le alternative inefficienti di modo da comprendere su quali variabili si debba intervenire per migliorare le prestazioni complessive.

Il *Chapter III* affronterà l'utilizzo di diverse tecniche di simulazione molecolare per la stima del coefficiente di ripartizione ottanolo-acqua (Kow), che è un proprietà fondamentale per il calcolo di alcuni indici utilizzati.

Il lettore troverà alcuni casi di studio descritti nel *Chapter IV*. Il primo appartiene al ramo della farmaceutica e si occupa della produzione del pioglitazone cloridrato attraverso l'utilizzo di diverse vie di sintesi appartenenti a numerosi brevetti. La seconda applicazione della metodologia riguarda l'industria biochimica e ottimizza le condizioni operative di un reattore utilizzato per la produzione di biodiesel a partire da olio vegetale. L'ultimo caso di studio esplora il mondo dei nanomateriali per applicazioni in ambito solare, valutando diversi parametri di reazione utilizzati per condurre la sintesi di un semiconduttore nanostrutturato, ovvero i quantum dot di seleniuro di cadmio (CdSe).

L'ultimo *Chapter V* conterrà le valutazioni conclusive e le prospettive future.

TABLE OF CONTENTS

ACRONYMS

AHP: Analytical Process Hierarchy

AIC: Akaike's Information Criterion

ANN: Artificial neural network

API: Active Pharmaceutical Ingredient

BAF: Bioaccumulation Factor

BCF: Bioconcentration Factor

BLIC: Battery-Limits Installed Cost

CAPE-OPEN: Computer-Aided Process Engineering Open

CAS: Chemical Abstract Service

CapEx: Capital Expenditures

CBMC: Configurational-bias Monte Carlo

CFs: Characterization Factors

COSMO: COnductor like Screening MOdel

COSMO-RS: COnductor-like Screening MOdel for Real Solvents

CP: Cost of the Project

CRS: Constant Returns to Scale

CSM: Continuum Solvation Model

DALY: Disability Adjusted Life Years

DEA: Data Envelopment Analysis

DfE: Design for Environment

DFT: Density Functional Theory

DMU: Decision Making Unit

ETs: Expression Trees

FAME: Fatty Acid Methyl Esters

FEP: Free Energy Perturbation

FOB: Free on Board

FTB: Freshwater Toxicity Potential

FWHM: full width at half maximum

GA: Genetic Algorithm

GEMC: Gibbs Ensemble Monte Carlo

GEP: Gene Expression Programming

GREENSCOPE: Gauging Reaction Effectiveness for the ENvironmental Sustainability of Chemistries with a multi-Objective Process Evaluator

GSCM: Green Supply Chain Management

H-Phrases: Hazard Phrases

HOMO: Highest Occupied Molecular Orbital

HTP: Human Toxicity Potential

ILCD: International Reference Life Cycle Data System

JRC: Joint Research Centre

Koc: soil sorption coefficient

kOH: reaction rate with hydroxyl

Kow: Octanol-Water partition coefficient

LabEx: Labor Expenditures

LCA: Life Cycle Assessment

LCI: Life Cycle Inventory

LCIA: Life Cycle Impact Assessment

LLR: Local Lazy Regression

LP: Linear Programming

LUMO: Lowest Unoccupied Molecular Orbital

m-SAS: modular-based Sustainability Assessment and Selection

MBEI: Material Balance Environmental Index

MD: Molecular Dynamics

MLR: Multiple Linear Regression

MOOH: Molecular Orbital OH

MSDSs: Material Safety Data Sheets

NPV: Net present Value

OA: oleic acid

ODE: 1-octadecene

OECD: Organization for Economic Co-operation and Development

OpEx: Operating Expenditures

OTA: Office of Technology Assessment

PAF: Potentially Affected Fraction

PBT: Persistent, Bioacculumative and Toxic

PCOP: PhotoChemical Oxidation Potential

PCR: Potential Chemical Risk

PDCA: Plan-Do-Control-Act

PDF: Potentially Disappeared Fraction

PEI: Potential Environmental Impact

PI: Profit Intensity

PLSR: Partial Least Square Regression

PPARγ: Peroxisome Proliferator-Activated Receptor Gamma

PPR: Project Pursuit Regression

PSD: Particle Size Distribution

QDs: Quantum Dots

QM: Quantum Mechanics

QSAR: Quantitative Structure-Activity Relationship

QSPR: Quantitative Structure Property Relationship

R-Phrases: Risk Phrases

REACH: Registration, Evaluation, Authorization and Restriction of Chemicals

R&C: Recommended but to be applied with caution

R&I: Recommended but in need of some improvements

R&S: Recommended and satisfactory

Rprop: Resilient backpropagation

SAR: Structure Activity Relationship

SDGs: Sustainable Development Goals

SSCM: Sustainable Supply Chain Management

SV: Support Vector

TBL: triple bottom line

TI: Thermodynamic Integration

TOP: trioctylphosphine

TOPSe: trioctylphosphine selenide

TQEM: Total Quality Environmental Management

TZD: thiazolidinedione

UNCHE: United Nations Conference on Human Environment

VRS: Variable Returns to Scale

WAR: Waste Reduction Algorithm

WCED: World Commission on Environment and Development

LIST OF FIGURES

LIST OF TABLES

CHAPTER I

SUSTAINABILITY AND SUSTAINABLE DEVELOPMENT:

AN OVERVIEW

I.1 SUSTAINABILITY AND SUSTAINABLE DEVELOPMENT CONCEPTS

The overall conditions of our planet Earth are progressively worsening.

Since we started to exploit the available natural resources, we have become one of the major sources of disturbance in a planet ruled by a multitude of non-equilibrium processes interacting in dynamic system. As soon as our major role in the climate change became clear, the scientific community strived for increasing the awareness of detrimental habits in the globalized society, aiming to slow down their effects on the environment, thus avoiding the reaching of boundaries of no return.[1–3] Indeed, the technology of modern society, developed to improve life quality and provide a higher satisfaction of human needs, e.g. better hygienic conditions, food stock granted by intensive farming and livestock, effective drugs, etc., generated evident benefits as well as dreadful drawbacks as they caused a considerable number of environmental issues that affect globally,[4] e.g. global warming,[5] ozone layer depletion,[5] air pollution,[6] land overexploitation,[7] or reduction of fresh water reserve[8] among the others.

In this context, an increasing effort has been spent in the last three decades on a smarter utilization of natural resources and fossil fuels, that will undeniably run out in the near future,[9] since an increasing adoption of renewable energy sources and a trend towards a reduction of the impact of human activities on the environment become essential in a sustainable development viewpoint.[10,11] Before going further, a brief description of "sustainable development" concept is necessary to deeper understand the implications of its adoption.

The term "sustainable" is generally defined in Cambridge Dictionary[12] as "*able to continue over a period of time*", embodying the concepts of perpetuity and conservation of a specific condition previously defined.

Conversely, the term "development" stands for "*the process in which someone or something grows or changes and becomes more advanced*",[12] providing an idea of improvement achieved through a progressive movement and evolution, that is an oxymoron in comparison to the static behavior related to the "sustainable" term.

In the particular context of socio-environmental systems, after few mentions regarding forestry in 17th century,[13] the first introduction of the modern concept of sustainable development has been conferred upon the UN Conference on Human Environment (UNCHE), held in Stockholm in 1972, while the adoption of the well-known actual terms has been attributed to the International Union for Conservation of Nature, that promulgate a World conservation strategy document[14] in 1980. The first comprehensive definition of the concept has been published in the Brundtland Report of 1987 by the World Commission on Environment and Development (WCED), being the sustainable development "*an enhancement process in which the exploitation of resources, the direction of investments, the orientation of technological development and institutional changes are made consistent with future as well as present needs."*[15] This definition comprehends two fundamental concepts: the "needs", seen as the essential living

needs to be fulfilled for the global increasing population focusing on the poorest, and the introduction of a threshold over which the environment is not resilient enough to absorb the effects induced by human activities and is not capable to sustain the actual social habits.[15]

Since the first definition of sustainable development concept, an abundance of new interpretations have been published, adapted to expectations specific for a desirable progress, as stated by Kates et al.: *"the concrete challenges of sustainable development are at least as heterogeneous and complex as the diversity of human societies and natural ecosystems around the world".*[16] The plethora of opinions and descriptions from scientists belonging to various disciplines gave rise to misinterpretations as none of them was capable of capture the whole spectrum of the concept.[17] Indeed, Johnston et al.[18] estimated that in 2007 there were around three hundred definitions of "sustainability" and "sustainable development". To cite but a few among many others, Mosovsky et al. described sustainability as "*the delivery of competitively priced goods and services that satisfy human needs and bring quality of life, while progressively reducing ecological impact and resource intensity throughout the life cycle, to a level at least in line with Earth's carrying capacity*";[19] Shaker R.R. defined "sustainable" as the peculiarity of the human-targeting homeostatic equilibrium between humanity and the surrounding ecological system, being the combination "sustainable development" the holistic approach and time-related processes that lead us to the end point of sustainability;[20] for Bolis et al. "*sustainable development can be seen as the kind of development aimed at satisfying the human needs of society as a whole (including future generations) beyond a minimum level, which is enabled by an axiological perspective in decision-making, considering environmental limits.";*[21] moreover García-Serna et al. asserted that "*sustainable development means continuous ensuring dignified living conditions with regard to human rights by creating, expanding, enlarging, refining and maintaining the widest possible range of options for freely defining life plans. The principle of fairness among and between present and future generations should be taken into account in the use of environmental, economic and social resources. Comprehensive protection of biodiversity is required in terms of ecosystem, species and genetic diversity and all of which the vital foundations of life are."*[22]

It is worth underlying how the definition of sustainable development should be globally acknowledged, since a lack of agreement on such a fundamental concept would lead to an imbalance in the evolutionary process among the different entities involved, compromising the future of the generations to come.[23]

Notwithstanding, the reader can easily recognize how they all harmonize with a common background, incorporating almost every aspect of the environment and human interaction therein,[24] and how they all share a common trend, that is including economic, social and environmental concerns simultaneously, although their time scales may differ, following the so-called triple bottom line (TBL). This concept has been proposed in 1997 by Elkington,[25] who introduced the three pillars of sustainability, i.e. people, profit and planet, that are still well-established benchmarks in the sustainable development discipline. After the ONG Earth Summit of 2002 held in Johannesburg, TBL has been referred to as the

balanced integration of economic, environmental and social performance.[26] A brief definition of these concepts is given below:

- "Economy" is the large set of inter-related production and consumption activities that aid in determining how scarce resources are allocated. This is also known as an economic system. The economy encompasses all activity related to production, consumption and trade of goods and services in an area. The economy applies to everyone from individuals to entities such as corporations and governments. The economy of a particular region or country is governed by its culture, laws, history, and geography, among other factors, and it evolves due to necessity. For this reason, no two economies are the same.[27]
- "Environment" stands for the complex of physical, chemical, and biotic factors (such as climate, soil, and living things) that act upon an organism or an ecological community and ultimately determine its form and survival.[28]
- "Society" is a group of people involved in persistent social interaction, or a large social group sharing the same geographical or social territory, typically subject to the same political authority and dominant cultural expectations. Societies are characterized by patterns of relationships between individuals who share a distinctive culture and institutions; a given society may be described as the sum total of such relationships among its constituent of members.[29]

The interactions among these entities have been represented in two different manners, which are reported in Figure I-1.

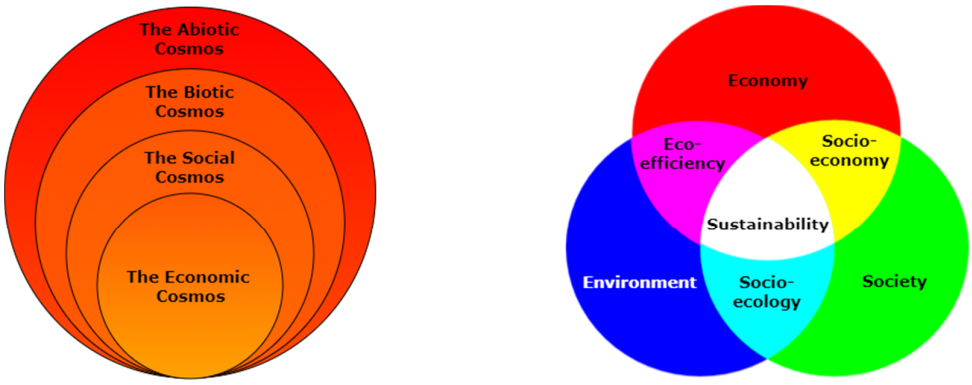

Figure I-1. Illustrations of the different approaches for representing the relationships among sustainability pillars. a) The Cosmic Model; b) The Dominant Model

The Cosmic Model (Figure I-1.a) has been postulated by Mebratu[17] under the following assumptions:

- the natural universe embeds (in the past and in the future) the human universe, in general, and the economic and social cosmos, in particular;
- the intersection area of the four cosmos is the area where we have the biggest amount of combinations of conflict and harmony that generates the opportunities for the development of the different scenarios of the natural and human universe;
- the interactions within the interactive zone are an abundance of systems that do not belong exclusively to the internal cosmo but have a four-dimensional (or three-dimensional, if we put the biotic and abiotic under the ecological dimension) systemic parameter;
- the environmental crisis recorded throughout human history is an outcome of the cumulative effect of deliberate, or otherwise, human neglect of one or more of the systemic parameters, resulting in millions of feedback deficient systems;
- not the entire abiotic region is influenced by the biotic, social and economic ones, same as a portion of the biotic cosmo, which is still independent from the human impact.

The Dominant Model (Figure I-1.b) has been proposed by Holmberg[30] and is based on the following statements:

- the system is reductionist: the environmental, economic, and social systems are independent entities and may be treated separately;
- the system is bivalent: sustainability is achieved in the interactive zone where the three different systems overlaps, while the area outside the interactive zone is assumed to be an area of contradiction;
- the system promotes linear thinking: since the ultimate objective of sustainability is the full integration of the natural, economic, and social systems; this may be achieved through the integration of these objectives.

In this work, the Dominant Model has been adopted.

The next *Paragraph I.2* will focus on the practical employment of sustainable development concept to the development of our society.

I.2 SUSTAINABLE DEVELOPMENT APPLICATIONS

The sustainable development approach has been applied by various entities on their own specific area of expertise. Since this manuscript deals with the sustainability assessment of chemical process plants, a particular concern will be given to chemical industry specific methodologies. Hence, the reader will find a summary of some of the application promoted by government institutions, then from industrial stakeholders and, last but not least, chemical industry practitioners and scientists.

I.2.1 SUSTAINABLE DEVELOPMENT AND INSTITUTIONS

The chance to put 1987 WCED definition[15] into practice showed up just five years later at the 1992 Earth Summit in Rio de Janeiro, when an action plan, called Agenda 21,[31] for the sustainable development of the 21th century implementation has been drawn, claiming for a broad cooperation among individuals in order to reshape the business model towards an integration of environmental and social concerns. Furthermore, Elkington[25] proposed a guideline of seven paradigms to be followed by enterprises in order to embed sustainability within their internal development, i.e. market (from submission to competitiveness), values (from rigid to flexible), transparency (from closed to open), life cycle technology (from product to function), partnership (from disunion to symbiosis), time conception (from intense to long term), and governance (from exclusive to inclusive).[21] In this perspective, Elkington advocated an involvement of organizations in the transition to a sustainable planet, since they ought to be concerned in social and environmental issues as well as economical ones.[25] Indeed, the European Union earmarked substantial funds (nearly €80 billion between 2014-20)[32] in a framework programme called Horizon 2020, funding research, technological development, and innovation. Horizon 2020 is turning into reality the European environmental research and innovation policy, providing a transformative agenda for promoting the inclusion of ecological concerns into economy and society as a whole, so as to achieve a genuine sustainable development: it is expected that at least 60% of the overall Horizon 2020 budget should be related to sustainable development, and that climate-related expenditure should exceed 35% of the budget.[33] In 2015, the United Nations General Assembly formally adopted the "universal, integrated and transformative" 2030 Agenda for Sustainable Development, employing a set of 17 Sustainable Development Goals (SDGs), that range from fighting poverty, hunger, discriminations and illiteracy to promoting clean energy, climate and natural life safeguard, good-health, sanitation, and recycling. Furthermore, international treaties have been signed, e.g. Kyoto Protocol in 1997 and Paris Agreement in 2015, by numerous countries (unfortunately, some of them are still missing or retreating) in order to ratify their commitment in the reduction of greenhouse gas production, responsible of global warming.

I.2.2 SUSTAINABLE DEVELOPMENT AND INDUSTRIAL SECTOR

As a matter of fact, we are slowly shifting from a reckless development towards a sustainable one and industry sector needs to be one of the driving forces of this process, since it has always played a major role in climate change.[34] Innovative products are constantly entering the market, answering to a growing demand from a wider global market, encouraging companies to invest in R&D more than ever before, even if process design still needs a big boost to gain compelling improvements.[35] Indeed, the US Office of Technology Assessment (OTA) defined "product design" (the specification of ''form and function'') as *"a unique point of leverage from which to address environmental problems. Design is the stage where decisions are made regarding the types of resources and processes to be used, and these ultimately determine the characteristics of waste streams*".[36] This definition means that an accurate product and process design would lead to clear environmental benefits, even though distribution, marketing, promotion, and pricing need to be accounted and analyzed in the future, as they have their own impacts on the final sustainability performance.[37] Moreover, within industries there are large gaps between proactive corporations, that are focused on preventing the generation of unsustainable development scenarios, and reactive companies, which internalize institutional regulations modifying their production processes, aiming to merely follow law perscriptions.[35] The improvements achieved during this enhancement process have to be identified in order to evaluate the order of magnitude of the increasing sustainability attitude. However, a relevant question arise at this stage. How can we measure the development achieved? Moreover, is it possible to quantify its sustainable pattern?

An abundance of sustainability metrics, methods, indicators and tools for the sustainability assessment have been proposed, aiming to quantify the performances of economic, social and environmental impacts of different systems, including transports, manufacturing, construction, supply chain and energy systems.[38,39] The scientific community agrees that the adoption of a single evaluation metric is not the desirable target to pursue, as dedicated methodologies would be more efficient when applied within their specific field,[40] even if it looks more profitable to consolidate all the available metrics into one aggregate system of metrics.[41]

Due to its central role, a considerable number of methodologies have been suggested for the sustainability evaluation of the industrial manufacturing sector[42] (which includes chemical, metallurgic, pharmaceutical, food, automotive, mechanical, textile, etc.) comprehending the total life cycle of the product, i.e. extraction of raw materials, transportation to manufacturing site, processing to obtain desired product, utilization by the end user and disposal of wastes, accounting for environmental, ethical and financial perspectives. The background that leads to the copious development of such an abundance of methodologies, finds its fundamentals on various models that took hold on the industrial mindset, i.e. Design for Environment, Life Cycle Analysis, Total Quality Environmental Management, Green Supply Chain Management, and ISO 14000:[43]

- *Design for Environment (DfE)*:[44] starting from the product design stage, it takes into consideration the potential refurbishing and/or recycle of the final product or some of its components, embedding their long term environmental and human impacts. Components are designed to exhibit interchangeability for reusable ones or biodegradability for consumables ones. Moreover, the raw materials extraction and the production processes efficiency, energy usage, water usage, and waste generated are analyzed, looking for their minimization.

- *Life Cycle Analysis*, also known as *Life Cycle Assessment (LCA)*:[45] it comprehends the potential environmental impacts and the resources exploited throughout the entire product's life-cycle from cradle, i.e. extraction of raw materials, to grave, i.e. waste disposal. The LCA methodology embodies four subsequent steps to perform a full-scale product assessment: Goal and Scope Definition, in which the target of the analysis has to be defined; Life Cycle Inventory (LCI), in which the information about mass and energy balances of materials, equipment, transformation process, transportation and waste treatment need to be retrieved; Life Cycle Impact Assessment (LCIA), which adopts the LCI information to calculate the impact of the entire life-cycle of the product based on impacts stored in databases; and Interpretation, in which the results from LCIA have to be analyzed to recognize the most sustainable pathway. The procedure of LCA are part of the ISO 14000 standard, in particular of ISO 14040:2006 and 14044:2006.

- *Total Quality Environmental Management(TQEM)*:[46] it is a high level framework adopted by companies to improve their environmental performance through a top down approach from management support through increasing employees' and stakeholders' awareness on environmental protection, integrating these policies in company's standard procedures, and measuring and recording environmental statistics, aiming to improve corporate's performance.

- *Green Supply Chain Management (GSCM)*: means integrating environmental thinking into supply-chain management, including product design, material sourcing and selection, manufacturing processes, delivery of the final product to the consumers as well as end-of-life management of the product after its useful life.[47] An extension of this concept is *Sustainable Supply Chain Management(SSCM)*, as it comprehends economic and social aspect in its concerns as defined by Ahi et al., who described SSCM as "*the creation of coordinated supply chains through the voluntary integration of economic, environmental, and social considerations with key inter-organizational business systems designed to efficiently and effectively manage the material, information, and capital flows associated with the procurement, production, and distribution of products or services in order to meet*

stakeholder requirements and improve the profitability, competitiveness, and resilience of the organization over the short- and long-term."[48]

- *ISO 14000 family standards*: provide practical tools for companies and organizations of all kinds looking to manage their environmental responsibilities.[49] The updated version of ISO 14001 is ISO 14001:2015 and it is based on the Plan-Do-Control-Act (PDCA) cycle to constantly improve the environmental performance of the manufacturing process. In the Plan phase, the company needs to review the objectives pursued and the processes required, the Do phase starts as soon as the processes are implemented, while in the Control step the company needs to check the environmental performances achieved, followed by the Act stage in which take place the evaluation of the results and the implementation of corrections aiming to improve the performances of the environmental management system. As already mentioned, LCA has been embedded within this standards in 2006.

I.2.3 SUSTAINABLE DEVELOPMENT AND CHEMICAL PROCESS ENGINEERING

Focusing on chemical industry, since its potential impact on the environment is a well-known issue,[50] more than a single guideline have been published, aiming to advise the practitioners on the improvement to implement within their chemical plants or during a process design phase. The sustainability analysis of the chemistry behind the processes has been performed by Anastas et al., whose Green Chemistry principles[51] are a benchmark in the area of chemistry and chemical engineering, as they provide a valuable endorsement for the minimization of the consumption of nonrenewable resources and the generation of hazardous substances. The twelve principles of Green Chemistry are reported underneath:

- it is better to prevent waste than to treat or clean up waste after it is formed;
- synthetic methods should be designed to maximize the incorporation of all materials used in the process into the final product;
- wherever practicable, synthetic methodologies should be designed to use and generate substances that possess little or no toxicity to human health and the environment;
- chemical products should be designed to preserve efficacy of function while reducing toxicity;

- the use of auxiliary substances (e.g. solvents, separation agents, etc.) should be made unnecessary wherever possible and innocuous when used;
- energy requirements should be recognized for their environmental and economic impacts and should be minimized. Synthetic methods should be conducted at ambient temperature and pressure;
- a raw material or feedstock should be renewable rather than depleting wherever technically and economically practicable;
- unnecessary derivatization (blocking group, protection/deprotection, temporary modification) should be avoided whenever possible;
- catalytic reagents (as selective as possible) are superior to stoichiometric reagents;
- chemical products should be designed so that at the end of their function they do not persist in the environment and break down into innocuous degradation products;
- analytical methodologies need to be further developed to allow for real-time, in-process monitoring and control prior to the formation of hazardous substances;
- substances and the form of a substance used in a chemical process should be chosen to minimize potential for chemical accidents, including releases, explosions, and fires.

Furthermore, the author supported the cause of Green Chemistry providing a specific framework, called Green Engineering,[52] conceived for scientists and engineers to engage in when designing new materials, products, processes, and systems that are benign to human health and the environment.[10] The twelve principles of Green Engineering follow:

- designers need to strive to ensure that all material and energy inputs and outputs are as inherently non-hazardous as possible;
- it is better to prevent waste than to treat or clean up waste after it is formed;
- separation and purification operations should be designed to minimize energy consumption and materials use;
- products, processes, and systems should be designed to maximize mass, energy, space, and time efficiency;
- products, processes, and systems should be “output pulled” rather than “input pushed” through the use of energy and materials;
- embedded entropy and complexity must be viewed as an investment when making design choices on recycle, reuse, or beneficial disposition;

- targeted durability, not immortality, should be a design goal;
- design for unnecessary capacity or capability (e.g., "one size fits all") solutions should be considered a design flaw.
- material diversity in multicomponent products should be minimized to promote disassembly and value retention;
- design of products, processes, and systems must include integration and interconnectivity with available energy and materials flows;
- products, processes, and systems should be designed for performance in a commercial "afterlife";
- material and energy inputs should be renewable rather than depleting.

In order to increase the control and awareness on the compounds available in the European market, the European Union promulgated a regulation, called REACH (Registration, Evaluation, Authorization and Restriction of Chemicals),[53] determining the standards and procedures regarding the production and use of chemical substances, and their potential impacts on both human health and the environment.

The contributions just reported, established the background over which, in the last two decades, scientists built up various methodologies to include sustainability in process system engineering and design. In the next *Paragraph I.2.3.1*, the reader will find a brief overview of the most used and widespread methodologies in chemical process design.

I.2.3.1 CHEMICAL PROCESS DESIGN METHODOLOGIES

It is well-known how a fair management and a responsible conduction of chemical process plants are crucial when environmental protection comes into play.[54] In this perspective, chemical engineers made a huge effort with the purpose of including the environmental and social evaluations beside the well-established economical one. Indeed, Azapagic et al.[55] first introduced a set of broad sustainability indicators already in 2000, embracing completely the establishing sustainable approach. Since then, an abundance of chemical and process industry tools, frameworks and methodologies[42] have been published in order to address the need of decision maker to choose the most sustainable design among a plethora of different alternatives. One of the most useful and widespread methodology is LCA, which has been already described in the previous paragraph.

Despite the clear benefits of this tool,[56] a full-scale LCA evaluation often requires reliable data, a high level of expertise as well as a considerable period of time to be performed for a lot of process designs,[57] particularly the LCI and LCIA steps. Hence, various software have been launched into the

market, answering to the urgency of user-friendly platforms to perform LCA calculations,[58–60] each one performing the assessment using customer input data or predefined databases, which could have been developed in-house or retrieved from external sources, e.g. ecoinvent[61] (adopted both for LCI and LCIA).[62]

A great number of contributions concerning LCIA methodologies can be retrieved in literature, covering various different impact categories, but providing characterization factors (CFs) that often differ among each other, even for the same chemical and impact, which yields usually confusion among practitioners.[24] Furthermore, the existing literature models embrace different perspectives for the CFs adopted: midpoint and endpoint indicators. A midpoint indicator can be defined as a parameter in a cause-effect chain or network (environmental mechanism) for a particular impact category that is between the inventory data and the category endpoints.[63] For instance, global warming potentials, ozone depletion potentials, and photochemical ozone formation potentials belong to this category. Endpoints are calculated to reflect differences between stressors at an endpoint in a cause-effect chain and may be of direct relevance to society's understanding of the final effect.[63] Among others, this category comprehend Damage to Human Health (measured using Disability Adjusted Life Years per kilogram of substance emitted, DALY/kg), and Damage to ecosystem diversity (Potentially Disappeared Fraction of species annually, PDF m^3 year). Figure I-2 shows the connection between these entities.

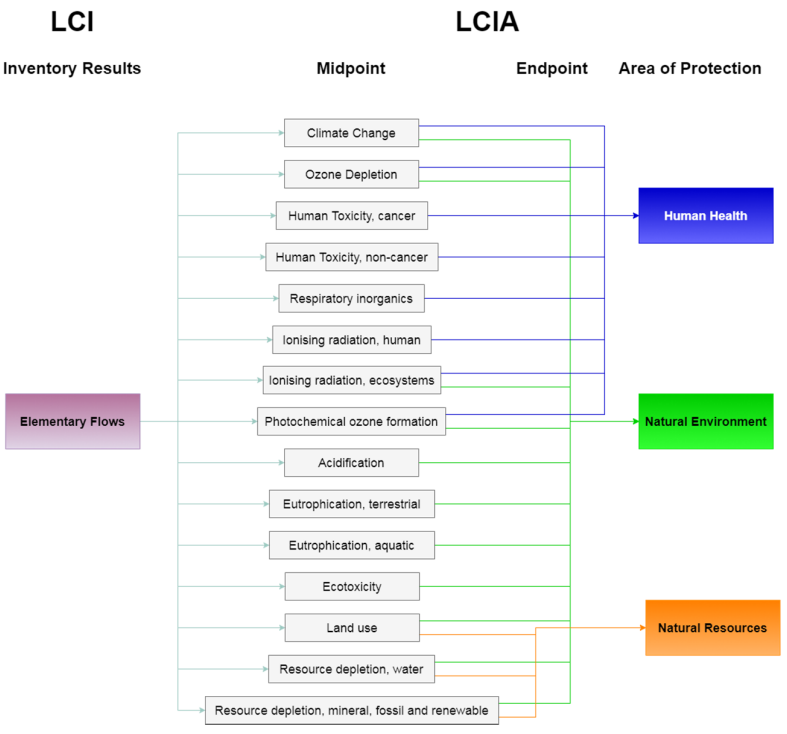

Figure I-2. Framework representing the connection between the results from LCI to 15 midpoint and endpoint indicators and the related 3 areas of protection (adapted from ILCD Handbook)[64]

With the purpose of identifying the best-practice among existing characterization models and providing recommendations to the LCA professionals, a review of the well-established methodologies in LCIA calculation has been provided by Hauschild et al., who worked in collaboration with the Joint Research Centre (JRC) of the European Commission that has launched the International Reference Life Cycle Data System (ILCD)[64] to develop technical guidance that complements the ISO Standards for LCA and provides the basis for greater consistency and quality of life cycle data, methods, and LCA studies.[65] In his research, Hauschild co-operated with experts and stakeholders of LCIA in order to select a set of relevant criteria, i.e. "completeness of scope", "environmental relevance", "scientific robustness and certainty", "documentation, transparency, and reproducibility", and "applicability", for screening the existing methodologies both at midpoint and endpoint. The LCIA methods included in the study are CML 2002,[66] Eco-indicator 99,[67] EDIP 2003,[68] EPS,[69] IMPACT 2002+,[70] LIME,[71] LUCAS,[72] ReCiPe,[73] and TRACI,[74] that combine various midpoint and/or endpoint indicators, but, although the numerous available combinations, none of them meets the international acceptance that the ISO standard calls for. Experts in LCIA assigned a level of recommendation to each CF considering the best existing techniques developed until 2008-2009 and the reliability of the final values obtained, as reported in Table I-1 which has been adapted from Hauschild et al.[65]. In Table I-1, "R&S" stands for "Recommended and satisfactory", "R&I" means "Recommended but in need of some improvements", "R&C" is the acronym for "Recommended but to be applied with caution" and "Interim" represents models not mature enough to be recommended.

Table I-1. Methodologies for calculation of characterization factors (adapted from Hauschild et al.)[65]

Impact Category	Midpoint		Endpoint	
	Best among existing characterization method	**Classification**	**Best among existing characterization method**	**Classification**
Climate change	Baseline model of 100 years of the IPCC[75]	R&S	Model developed for ReCiPe[73]	Interim
Ozone depletion	Steady-state ODPs from the WMO assessment[76]	R&S	Model for human health damage developed for ReCiPe[73]	Interim
Human toxicity, cancer effects	USEtox model[77]	R&I/R&C	DALY calculation applied to USEtox midpoint[78]	R&I /Interim
Human toxicity, non-cancer effects	USEtox model[77]	R&I /R&C	DALY calculation applied to USEtox midpoint[78]	Interim
Particulate matter/respiratory inorganics	as in Humbert[79] based on Rabl and Spadaro[80] and Greco et al[81]	R&S / R&I	Adapted DALY calculation applied to midpoint (adapted from van Zelm et al.,[82] Pope et al.[83])	R&S / R&I
Ionizing radiation, human health	modelled by Dreicer et al.[84] in Frischknecht et al.[85]	R&I	Frischknecht et al.[85]	Interim
Ionizing radiation, ecosystems	Screening Level Ecological Risk Assessment[86] based on AMI model[87]	Interim	None identified	-
Photochemical ozone formation	LOTOS-EUROS as applied in ReCiPe[82]	R&I	Model for damage to human health as developed for ReCiPe[88]	R&I
Acidification	Accumulated exceedance[89,90]	R&I	Method developed by van Zelm et al.[91]as in ReCiPe[73]	Interim

Eutrophication, terrestrial	Accumulated exceedance[89,90]	R&I	No methods found	-
Eutrophication, aquatic	EUTREND model as implemented in ReCiPe[82]	R&I	Model for damage to ecosystem (freshwater only)[73]	Interim
Ecotoxicity, freshwater	USEtox model[77]	R&I/R&C	None identified	-
Land use	Model based on soil organic matter (SOM)[75]	R&C	Model for species diversity loss as in ReCiPe[73]	Interim
Resource depletion, water	Model for water consumption as in Swiss ecoscarcity,[92]CML 2002[66]	R&I	None identified	-
Resource depletion, mineral and fossil	CML 2002[66]	R&I	Resource depletion, mineral and fossil Method developed for ReCiPe[73]	Interim

R&S: Recommended and satisfactory; R&I: Recommended but in need of some improvements; R&C: Recommended but to be applied with caution; Interim: not mature enough to be recommended

As the reader can notice, a great number of different methodologies needs to be employed in order to encompass the wide variety of possible impacts on sustainability, resulting in a multi-tool time-demanding procedure. Since none of the CF is negligible, practitioners are aware that they should include the whole set of methods to gain reliable results from the LCIA analysis. Luckily, some of them have already been integrated within a comprehensive method, i.e. Recipe2016,[73] which can be adopted by itself or in combination with the available LCA calculation platforms.[58–60] However, sometimes practitioners prefer not to perform a full scale LCA due to some overall procedure intrinsic issues:

- when there are a multitude of different scenario and/or various synthetic routes to assess in order to obtain the same product, the effort endorsed in the full-scale LCA of the whole variety of different designs becomes unbearable, as each alternative requires a considerable amount of time to be analyzed;
- when the substances involved within the chemical processes are not included in LCA databases, an extensive literature review becomes necessary in order to fill the data gaps in compound properties. During this process, it is not uncommon to pinpoint a lack of literature data for some chemicals, which prevents the accomplishment of the evaluation or add uncertainty to the final result;
- when the company can't afford to allocate a substantial amount of resources on R&D, thus a series of test to retrieve experimental data for LCA is not feasible;
- when entering the market with an innovative product guarantees a great advantage on the competitors, therefore the company needs to complete the LCA evaluation as soon as possible to achieve a shorter time-to-market.

These scenarios are not uncommon for chemical industry corporates devoted to innovation like the ones belonging to the pharmaceutical field, that is characterized by a plethora of different process

alternatives for the same active pharmaceutical ingredient (API), dearth of published data on the chemicals adopted, R&D departments usually overloaded by a multitude of experiments and more focused on the effects of the final API, and a substantial interest to patent APIs before the competitors in order to gain a predominant position in the market.[93]

In this context, pharmaceutical companies are used to consider fewer CFs or adopt methodologies that comprehend a minor number of contributions, trying to find the best trade-off between the quickness and the amplitude of the sustainability assessment. This is acceptable provided that the three pillars of sustainability are accounted and a full-scale LCA is performed on the most efficient designs, after the preliminary screening of the inefficient ones.

With the purpose of providing valuable tools to assist practitioners during a short process design assessments, plentiful methods have been published in the last decade, trying to couple process modelling, generation of process alternatives, sustainability evaluation and retrofit analysis. Just to cite a few, Gonzales et al.[94] proposed a framework called Gauging Reaction Effectiveness for the ENvironmental Sustainability of Chemistries with a multi-Objective Process Evaluator (GREENSCOPE), aiming to estimate process alternative sustainability on a wide perspective, comprehending indicators accounting for different concerns, i.e. Efficiency, Energy, Economics, and Environment, while ensuring an interdependence among them. GREENSCOPE proved to be a versatile tool, since the battery limit of the evaluation can vary, from unit operation to overall process plant, and several raw materials, synthesis path, manufacturing technology and byproducts can be accounted for the same final product. Lapkin et al.[95] developed a methodology organized on four hierarchical levels, i.e. society, infrastructure, company, and product and process, for the evaluation of the sustainability, called "greenness", of different available technologies, where each level is distinguished by a set of suggested established indicators based on stakeholders' relevance. Sikdar et al.[96] presented a framework composed by a set of four indicators that accounted for the three dimensions (3D) of sustainability. Two of them were focused on process operation, while the other two on human health risk and environmental impacts related to the chemicals adopted. This framework has been chosen by Fermeglia et al.[97], coupled with Waste Reduction Algorithm (WAR),[98] process simulation with a CAPE-OPEN standard and quantum-mechanics molecular modelling, for the implementation of Process Sustainability Prediction (PSP) Framework, aiming to assess the sustainability performance of chemical process designs. Othman et al.[99] published a modular-based Sustainability Assessment and Selection (m-SAS) framework which embeds Analytical Process Hierarchy (AHP) in order to guide practitioners during the selection of alternative process designs. m-SAS integrates four modules (process simulation, equipment and inventory acquisition, sustainability assessment, decision support) to assist the development of models for case studies, data acquisition and analysis, team contribution assessment and decision support process, respectively. Tugnoli et al.[100] proposed the Quantitative Assessment of Sustainability Indices method based on the quantitative calculation of a set of normalized impact indices, i.e. sixteen environmental, one economic and two social,

embedded in a four steps procedure, in which, after a selection of common reference criteria among alternative designs, the indicators are defined, normalized and aggregated into the final result. Torres et al.[101] published the Material Balance Environmental Index (MBEI) framework which allows to estimate toxic properties of the substances involved inside the chemical plant and, coupled with HYSIS simulation software used to determine mass flows, provide a sustainability evaluation using simple index calculation based on toxic properties formerly estimated. Shadiya et al.[102] built up an Excel based tool called SUSTAINABILITY EVALUATOR, with the purpose of assessing processes' sustainability, considering a selection of metrics that concern economic, environmental, and health and safety issues. SustainPro is a tool designed by Carvalho et al.[103] in order to guide the user throughout four steps during the design of continuous or batch processes. The first step requires the definition of process flowsheet specifications, the second one prescribes the calculation of indicator to identify potential bottlenecks, the third embeds safety, economic and environmental evaluations, while the last one accounts for designing possible alternatives, that are further evaluated using environmental impact tools and safety indices. Babi et al.[104] developed a three stages framework able to generate plausible superstructure networks for converting given raw materials to potential final products using available technologies. After the generation of the possible alternatives (synthesis-stage), the selection of the optimal design paths considering sustainability indicator follows (design-stage), concluding with the introduction of process intensification criteria for the generation and evaluation of alternatives that match the desired improvement targets (innovation-stage).

In this manuscript, a methodology for a preliminary screening of process alternatives has been introduced in order to reduce the number of full-scale LCA evaluation, excluding a priori sub-optimal designs. The adoption of the methodology proposed underneath will speed up the design stage, reducing the effort and the money involved in the definition of the process alternative to further develop, and will provide some solutions for data gaps filling using molecular modelling. It consists of five sequential steps as shown in Figure I-3:

1. retrieving different synthesis routes published in literature or provided by in-house experiments or know-how;
2. creating a flowsheet related to each possible design using process simulators in order to obtain mass and energy balances;
3. choosing and calculating appropriate well-established indicators for the sustainability evaluation among the ones proposed in literature according with data available, impact addressed and decision-maker concerns;
4. applying Data Envelopment Analysis (DEA) to identify the most sustainable among alternatives;

5. performing a retrofit analysis, identifying the major contributions on sub-optimal designs indicators and providing a qualitative feedback regarding the parameters that should be modified (if possible) to improve the scores of each design.

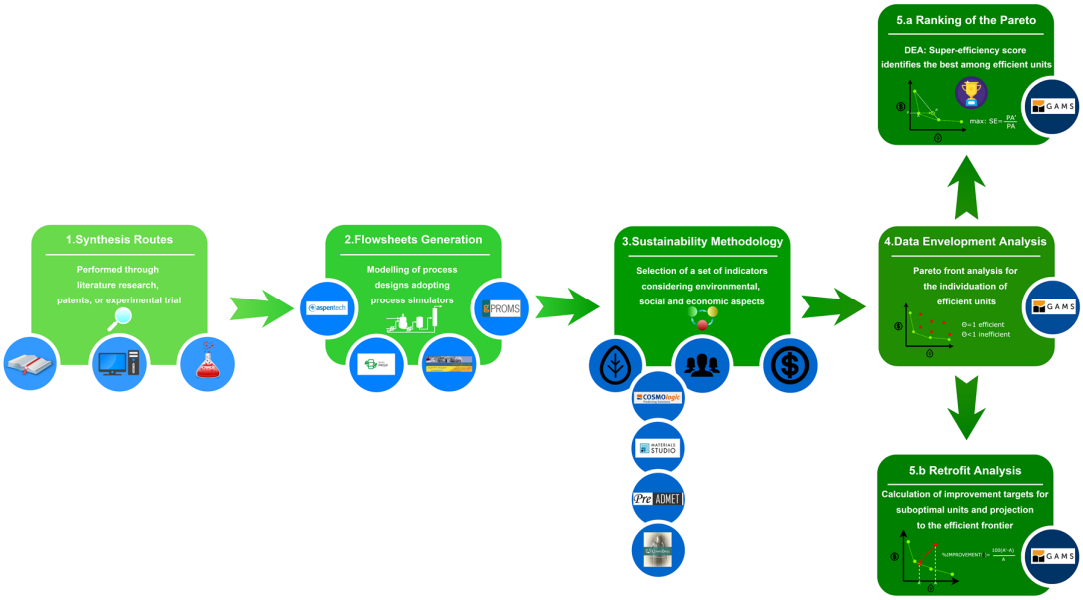

Figure I-3. Methodology qualitative decision path.

In the next *Chapter II* an extensive description of each step will be presented, including the characteristics of the indicators chosen as well as the fundamentals of DEA.

CHAPTER II

SUSTAINABILITY EVALUATION METHODOLOGY

One of the most valuable parameter to consider during a process design phase is time. It is crucial for a company to speed up as much as possible the launch of a product into the market in order to reduce the R&D costs and take an advantage on the competitors. In this perspective, the methodology proposed in this manuscript fulfills the need of avoiding the waste of time caused by performing extensive analysis on inefficient process designs, although an extensive LCA assessment on the efficient designs using established procedures is still required afterwards. The reader will find underneath a description of the five main steps to adopt in order to perform the methodology proposed.

II.1 RETRIEVING DIFFERENT ROUTES

In order to enter the market with an established product, a company needs to review the existing literature to identify as many synthesis routes as possible, therefore this step requires the access to patents, literature databases and books. The collection of documents needs to be followed by a check on the expiration date of patents and the replication of routes among different sources. Some trivial changes among different designs are sometimes not worth to assess, whereas a single paper can provide more than one promising route. During this step, a first guess on equipment employed and missing data will result in a homogeneous approach for all designs under investigation, ensuring to avoid imbalance which could end up with misleading evaluations.

II.2 MODELING ROUTES ADOPTING PROCESS SIMULATORS

Synthesis routes have to be modelled in a process simulator in order to obtain energy and mass balances. Process simulators are a well-established tools adopted to design, develop, analyze and optimize chemical processes.[105] In this design phase, the company will take advantage of its know-how about scaling-up, phase equilibria and thermodynamic properties of the compounds involved in the processes. Some quantitative or qualitative data can be usually retrieved on patents, e.g. reaction yields and selectivity, byproducts, extraction of solutes, operating conditions, phase separation performances and equipment sizing, therefore it is possible to include this information within process simulations in order to achieve a reliable model. In the event that simplifications needs to be assumed, a similar approach on approximations for each design should be adopted to prevent generation of biases. Flowsheets should be scaled up to the actual demand of desired product in order to evaluate the energy consumption of solvent recovery, while an optimization of the scheduling should be performed for batch process designs.

After the completion of the steps described in *Paragraphs II.1* and *II.2*, practitioners concluded a task equivalent to a shortened estimated LCI.

II.3 CHOOSING AND CALCULATING INDICATORS

It is well known how economy, society and environment are connected and dependent upon each other in a sustainable development viewpoint, thus it is essential to determine an adequate trade-off among them. The selection of a suitable set of sustainability indicators is fundamental, even though the accuracy of the results of the assessment is particularly related to trustworthy experimental or literature data. In order to preserve the simplicity of this approach the indicators endorsed should not require too many information to be calculated or, alternatively, the data should be assessable using computational techniques. Since this methodology can be performed using any set of indicators, for illustrative purposes various well-established indicators have been adopted in this work, aiming to investigate the capabilities of different approaches. The information retrieved for the quantification of the indicators have been mass and energy balances of the process, equipment costs from vendors or process simulators, H-Phrases, substance structures and chemicals costs from literature, and several thermodynamic properties, including Kow (Octanol-water partition coefficient) and kOH (reaction rate with hydroxyl) from molecular modelling.

II.3.1 ECONOMIC PERFORMANCE EVALUATION

II.3.1.1 COST OF THE PROJECT (CP)

CP quantifies the economic performances of each process design accounting for both Capital Expenditures (CapEx) and Operating Expenditures (OpEx) as shown in Eq.1, which is adapted from Schaber[106] and evaluates the discounted total cost of the project.

$$CP = (CapEx) + \sum_{i=1}^{\tau} \frac{(OpEx)}{(1 + r_d)^i} \qquad Eq.1$$

in which r_d is the discount rate, while τ is the project lifetime.

The CapEx calculation is based on the elements reported in Table II-1, which is adapted from Schaber's work.[106] FOB (free on board) cost represents the cost of each process unit, ignoring additional related expenses, e.g. ancillary equipment, delivery, electrical, engineering, or piping. The delivery of each unit is accounted by increasing its FOB costs by 5%[106] and then multiplied using a Wroth factor,[107] which is an heuristics based coefficient adopted to include the additional expenses formerly excluded by FOB, obtaining the battery-limits installed cost (BLIC) of each piece of equipment. Furthermore, it is necessary to include some extra expenditures related to the production site, providing a reliable estimation of equipment unit cost based on heuristics. Therefore, every BLIC needs to be multiplied by some extra

coefficients related to each additional plant-related cost, as shown in Table II-1. Since this work is addressed mainly to pharmaceutical corporates, it is preferable to adopt the upper limit value[107] to coefficients from (5) to (8), due to the rigorous hygiene regulations and the small scale of productions typical of pharmaceutical field.[106]

Table II-1. CapEx Heuristics[107]

Item	***Cost***
(1) FOB cost	sum of processing equipment units, solvent recovery excluded
(2) delivery	5% of FOB cost
(3) installation: ancillary equipment, automation, electrical, piping, and engineering	[(Wroth Factor)-1]·(delivered equipment cost)
(4) battery-limits installed cost (BLIC)	sum of items (1) to (3)
(5) buildings and structures	10-20% of BLIC
(6) contingency	15-20% of BLIC
(7) offsite capital (for a grass-roots plant)	45-150% of BLIC
(8) service facilities	10-20% of BLIC
(9) waste disposal	not included in CapEx
(10) working capital	30-35% of annual materials costs
(11) total CapEx	sum of items (4) to (10)

OpEx estimation includes raw materials cost, utilities cost and labor expenditures (LabEx), as shown in Eq.2.

$$OpEx = \sum_i \frac{(m_i^{in} - k_i \cdot m_i^{out})}{m_{prod}} \cdot price_i + \sum_j \left(\frac{m_j^{in}}{m_{prod}} \cdot price_j \right) + LabEx \qquad Eq.2$$

in which m_i^{in} and m_i^{out} are, respectively, inlet and outlet mass of raw materials only, m_{prod} is the total mass of main product obtained, $price_i$ is the market price per unit of each chemical, m_j^{in} is the mass of utility j adopted at the $price_j$, and k_i takes into account the estimated percentage of recovery of solvents, unreacted reagents and catalysts (it can vary in the range between 0 and 1), in which 0 means a total utilization of a chemical i within the process and 1 indicates a total recovery of the substance for further utilization.

LabEx is evaluated considering the operators' annual gross salary times the number of operators employed. Both information are embedded in process simulators or can be estimated using well-established methodologies.[108]

CP calculation is not related to expected revenues as they depend on forecasted product market price and volumes. Therefore CP adoption is suggested whenever a reliable forecast on future revenues is unachievable, as it would increase the uncertainty of the sustainability analysis.

II.3.1.2 NET PRESENT VALUE (NPV)

The adoption of NPV is preferable in comparison with CP as soon as the market price of the product is well-known or a set of historical data is available for a trustworthy forecast of future revenues. The formula for quantification of NPV is reported underneath (Eq.3) and it differs from Eq.2 (CP calculation) by the last revenue dependent term.

$$\text{NPV} = -(CapEx) + \sum_{i=1}^{\tau} \left\{ \frac{-(OpEx)}{(1+r_d)^i} + \frac{revenues}{(1+r_d)^i} \right\} \qquad Eq.3$$

In which every term refers to the same entity as in Eq.2 and revenues are calculated considering the market price of the product times the mass produced annually.

II.3.1.3 PROFIT INTENSITY (PI)

In addition, it is possible to assess the economic performance of various process alternatives during a very early experimental phase, when neither the equipment nor the optimal quantities of compounds have been defined yet. Despite the limited information available at this stage, PI provides a valuable indication on future costs considering material balance only. It is based on Material Intensity proposed by Martins et al.,[96] which quantifies the amount of non-renewable sources required to obtain a unit mass of products, including raw materials and solvents. In order to evaluate the economic impact of the distinct synthesis routes, the cost of each substance has been embedded in the calculation path, giving a first estimation of future material costs (Eq.4 and Eq.5).

$$\text{PI} = \frac{\sum_i m_i^{in} \cdot price_i - \sum_j m_j^{out} \cdot price_j}{output} \qquad Eq.4$$

$$output = \frac{\sum_j m_j^{out}}{m_{prod}} \qquad Eq.5$$

where m_i^{in} is the total inlet massflow of raw material i times its purchasing price, $price_i$, m_j^{out} is the outlet massflow of the saleable product j which is multiplied by its selling price, $price_j$, and m_{prod} is the total outlet massflow of the desired product of the synthesis. In case a single saleable product is obtained, the term "output" equals unity, therefore Eq.4 is simplified as shown in Eq.6.

$$\text{PI} = \sum_i (m_i^{in} \cdot price_i) - m_{prod} \cdot price_{prod} \qquad Eq.6$$

II.3.2 SOCIAL PERFORMANCE EVALUATION

II.3.2.1 POTENTIAL CHEMICAL RISK (PCR)

Potential Chemical Risk (PCR) addresses hazards and risks for humans related to chemicals handled in the manufacturing process in order to exhibit the social contribution of a process design to sustainability. This indicator is based on the work of Vincent et al.,[109] in which risk classes were related to R-Phrases assigned to each chemical. Aiming to define the correct human risk category in which a substance relies, R-Phrases were classified in risk classes considering an increasing risk accordingly to the increasing classes. Nowadays, R-Phrases have been replaced by H-Phrases,[110] so that an update of risk classes to hazard classes using H-Phrases was required, as shown in Table II-2. In particular, hazard class 0 is specific for chemicals with a lack of data in literature, while hazard class 1 is reserved for well-known harmless compound.

Table II-2. PCR hazard classes

Hazard Class	***H-Phrases***
0	Unknown
1	None
2	H290, H303, H305, H313, H316, H317, H333, H334, H336, EUH066
3	H205, H221, H223, H226, H227, H228, H230, H231, H242, H251, H252, H261, H272, H302, H304, H312, H315, H319, H332, H335, H341, H351, H361, H362, H371, H373, EUH201, EUH202, EUH203, EUH204, EUH205, EUH209A

4	H204, H225, H240, H241, H250, H260, H270, H301, H304, H311, H314, H318, H331, H340, H350, H360, H370, H372, EUH001, EUH014, EUH018, EUH019, EUH070, EUH209
5	H200, H201, H202, H203, H220, H222, H224, H271, H300, H310, H330, EUH006, EUH071

The calculation of PCR needs mass balance from process simulator and H-Phrases related to each chemical and is made up of assignments of each chemical to specific hazard classes whose contribution define the overall indicator. Whenever more than one H-Phrase is assigned to a substance, the selected chemical hazard class is the highest possible among the classes related to its H-Phrases. The calculation of the indicator is shown in Eq.7 in which m_i^{max} is the maximum mass flow of component *i* (inlet flow for raw materials, catalysts and solvents, outlet flow for products and byproducts), m_{prod} is total mass of main product obtained in the process and $H_{Cl_i}^{PCR}$ is the hazard class assigned to chemical *i* and related to risks to human health. Since it is advisable to assess the maximum potential risk related to the adoption of a specific substance, users should account the maximum amount of each chemical i within the process plant.

$$PCR = \sum_i \frac{m_i^{max}}{m_{prod}} \cdot 10^{H_{Cl_i}^{PCR}} \qquad Eq.7$$

Despite Vincent's work,[109] this methodology do not consider frequency class, that accounted for the duration of the utilization of each substance, nor quantity classes of chemicals, which concerned the quantity of chemicals involved within the chemical plant in comparison with the most used one. The former has been neglected since a permanent adoption of every substance has been assumed, the latter has been replaced by the maximum mass of chemical per mass of product, since the adoption of quantity classes was leading to possible misinterpretation of the results. Indeed, the sorting of substances to quantity classes (from 1 to 5) was funded on the choice of ranges in which each chemical should have belonged accordingly to the value of the ratio of the amount of substance to the quantity of the most used chemical in the process. The ratios varied from values close to 0 for traces of a substance, to 1 for the most used chemical, therefore the quantity classes were defined in the range 0-1, i.e. (quantity class 1)<0.01, 0.01≤(quantity class 2)<0.05, 0.05≤(quantity class 3)<0.12, 0.12≤(quantity class 4)<0.33, (quantity class 5)≥0.33. This was leading to erroneous conclusion whenever a different number of substances was adopted among the various alternative designs. For instance, assume that a practitioner desires to assess two alternative designs, i.e. "Design 1" and "Design 2", for the production of the same final product using different solvents. "Design 1" adopts 10 kg of hazard class 5 "Solvent A", while "Design 2", adopts 5 kg of same "Solvent A" and 5 kg of hazard class 4 "Solvent B" which is less hazardous. Both

designs employ the same quantity of solvents (10 kg), while "Design 2" is improving its sustainability performance, reducing the amount of very hazardous "Solvent A", in favor of a less hazardous "Solvent B": this should produce a minor impact and a lower value of the PCR score, since, in Vincent's approach, the quantity class in combination with the hazard class provides the contribution to the final PCR value related to each chemical. Instead, the PCR calculation leads to an opposite outcome, since the quantity classes are based on the most used substance. Indeed, for "Design 1" there is only one substance belonging to quantity class 5 so there is just one contribution (which is 10^6 accordingly to Vincent's work)[109] to the final PCR value, while for "Design 2" there are two substances belonging to quantity class 5, so there are two contributions (specifically $10^6 + 10^5$, which are different considering their specific hazard classes) to the final PCR value, which is actually increasing the final PCR value instead of decreasing it. Moreover, the ratios of the other substances increase as the quantity of the most used substance decreases, causing a greater contributions from other chemicals employed in the process.

In order to avoid this kind of scenarios, quantity classes have been neglected in favor of the introduction of maximum mass of chemical per mass of product, which is preserving the ratios among the various chemicals and accounts properly the number of substances. Considering the example above, assuming that both designs produce the same amount (1 kg) of final product, the "Solvent A" contribution to "Design 1" is 10^6 (Eq.7), meanwhile the solvents of "Design 2" generate impacts for $5 \cdot 10^5$ and $5 \cdot 10^4$, respectively for "Solvent A" and "Solvent B", with a final contribution of $5.5 \cdot 10^5$ to PCR final value. The drop of the PCR final value is in accordance with Green Chemistry principle of reducing the hazardous chemical adopted, thus the latter approach is the one endorsed in this thesis.

Furthermore, the relationships among the different hazard classes contributions to final PCR score from Vincent's method[109] have been preserved. Indeed, each growth of one hazard class caused an increment of one order of magnitude to chemical contribution to the final PCR score. For instance, Vincent affirmed that two chemical belonging to the same quantity and frequency classes (let's assume 4 and 4) generate an impact quantified in 10^5 or 10^6, whether they belong to hazard classes 4 or 5, respectively. The relationships among classes have been retained in Eq.7, since the hazard class is the exponential term of a base 10, meaning that an increasing of one hazard class induces an order of magnitude growth.

II.3.2.2 HUMAN TOXICITY POTENTIAL (HTP)

This indicator, as PCR, accounts for human health risks related to the chemicals handled within the chemical plant in order to quantify the social concerns associated with the product manufacturing. The hazard classes related to this indicator are independent from H-Phrases, therefore its adoption is encouraged if the majority of the chemicals under study lacks of these information. The structure of the equation for quantifying HTP, i.e. Eq.8, is similar to the one adopted for the calculation of PCR (Eq.7) as shown below:

$$\text{HTP} = \sum_i \frac{m_i^{max}}{m_{prod}} \cdot H_{Cl_i}^{HTP} \qquad Eq.8$$

in which m_i^{max} represents, as stated before, the maximum mass flow of component *i* (inlet flow for raw materials, catalysts and solvents, outlet flow for products and byproducts), m_{prod} is total mass of the product obtained in the process and $H_{Cl}{}_i^{HTP}$ is the hazard class assigned to chemical i and related to risks to human health. The difference with Eq.7 resides on the calculation of the hazard class term, which is related to various endpoints estimated using literature available data or, more likely, in-silico methods. For instance, in this work the QSAR tool HazardExpert,[111] embedded in the software Pallas by CompuDrug,[112] has been adopted to achieve this task. HazardExpert is a rule-based software tool for predicting the toxicity of organic compounds in humans and in animals based on known toxic fragments collected from in vivo experiments and reported by the US EPA. This tool is able to link toxic molecule segments to their effects on various biological systems, combining the use of toxicological knowledge, expert judgement, QSAR models, and fuzzy logics (which simulates the effects of different exposure conditions). The fundamentals have been developed using fragment databases from different sources, whilst new fragments can be implemented by the user, providing an open architecture knowledge base. It comprehends multiple toxicity endpoints which are exploited to investigate human health hazard effects including oncogenicity, mutagenicity, teratogenicity, membrane-irritation, sensitization, immunotoxicity, and neurotoxicity. The probability of a toxic effect is calculated for any single endpoint, followed by an overall toxicity probability in the range from 0, which means that the compound is unlikely to be toxic, to 100, which gives a high probability of acute toxic effects. The overall probability has been chosen to represent the toxicity of a compound, assigning its value to $H_{Cl_i}^{HTP}$ for each substance i.

So far, $H_{Cl_i}^{HTP}$ was related to the probability of a compound of being toxic considering various endpoints simultaneously aggregated in a singular percentage. Depending on the toxicity evaluation output of the simulation tools available, the user has sometimes to handle a single output for each endpoint considered, as in the case of PreADMET.

The web-based application PreADMET,[113] based on backward elimination[114] and resilient backpropagation (Rprop)[114] neural network method,[115] has been adopted for the evaluation of specific toxicity assays, since it has already been used for the evaluation of toxicity elsewhere.[116] Virtual screenings of mutagenicity (estimated Ames test against strains of Salmonella typhimurium TA100 and TA1535),[117] rodents carcinogenicity bioassays (based on National Toxicology Program and FDA US data on in-vivo 2 year carcinogenicity tests of mice and rats) and risk of inhibition of human ether-a-go-go-related (hERG inhibition estimation)[114] gene have been performed to evaluate toxicological likelihood, starting from 2D structural models of chemicals. The results provided are “negativity” or “positivity” to mutagenicity and carcinogenicity (against mice or rats) assays and “low, medium or high risk” for hERG

inhibition. Therefore, as shown in Eq.9, a related assignment to H_{ji}^{HTP} of values 0 or 1, respectively for "negative" or "positive" event, and of 0, 0.5 or 1 respectively for low, medium of high inhibition risk has been endorsed.

$$H_{ji}^{HTP} = \begin{cases} 1 & if\ positive\ or\ hERG\ high\ risk \\ 0.5 & if\ hERG\ medium\ risk \\ 0 & if\ negative\ or\ hERG\ low\ risk \end{cases}$$

$$H_{Cl_i}^{HTP} = \sum_j H_{ji}^{HTP} \qquad Eq.9$$

where H_{ji}^{HTP} is the specific value (0, 0.5 or 1) assigned to chemical i regarding the biological endpoint j.

Since in this manuscript four biological assays have been considered, the value of $H_{Cl_i}^{HTP}$ varies in a range between 0 and 4, however practitioners are encouraged to adopt the biological assays that are more likely to satisfy their particular needs or the significant ones that can be estimated using the information available.

II.3.3 ENVIRONMENTAL PERFORMANCE EVALUATION

II.3.3.1 POTENTIAL ENVIRONMENTAL IMPACT (PEI)

This indicator quantifies several environmental contributions to sustainability of process alternative designs. It is based on mass balance, transfer coefficients and H-Phrases relying on Vincent's work,[109] although the hazard classes have been updated to replace R-Phrases with H-Phrases in Table II-3, using the same procedure as for PCR. It is worth underlying how H-Phrases are rather easy to retrieve in literature and, in case of lack of information about some compounds, an hazard class can still be assigned to compound considering the most optimistic scenario (hazard class 0) or the most pessimistic one (hazard class 5).

Table II-3. PEI hazard classes

Hazard Class	H-Phrases
0	Unknown
1	None
2	H303, H305, H313, H316, H320, H333, H336, H413, EUH066, EUH029 and accidental probability of a contact with water, EUH031 and accidental probability of a contact with acid
3	H251, H252, H261, H302, H304, H312, H315, H319, H332, H335, H341, H351, H361, H362, H371, H373, H402, H412, EUH201, EUH202, EUH203, EUH204, EUH205, EUH029 and occasional probability of a contact with water, EUH031 and occasional probability of a contact with acid, EUH032 and accidental probability of a contact with acid
4	H250, H260, H270, H301, H304, H311, H314, H318, H331, H340, H350, H360, H370, H372, H401, H411, H420, EUH001, EUH014, EUH018, EUH019, EUH070, EUH029 and permanent probability of a contact with water, EUH031 and permanent probability of a contact with acid, EUH032 and occasional probability of a contact with acid
5	H200, H201, H202, H203, H271, H300, H310, H330, H400, H410, H420, EUH032 and permanent probability of a contact with acid

The PEI calculation follows a decision path to define the total potential environmental impact, based on mass flows and H-Phrases of chemicals, as well as on the transfer coefficients. The last one is driven by physicochemical properties and is related to physical state of substances (such as gas, liquid, solid or powder) and the medium (such as air, water or soil) in which a release in the environment is more likely to happen (Figure II-1).

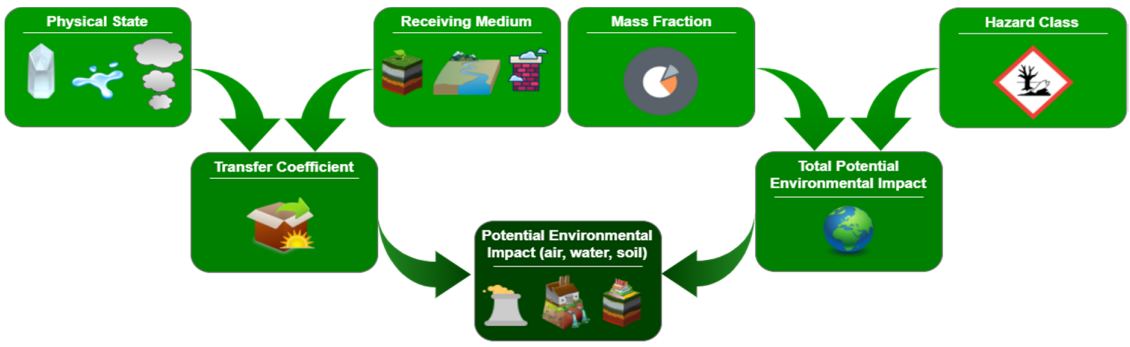

Figure II-1. PEI Algorithm

The introduction of a transfer coefficient provides a more realistic impact evaluation, as defined in Vincent's work[109], taking into account the relationship between the physical state of a compound and the media in which is more likely to be released. Transfer coefficients are shown in Table II-4, which has been adapted from Vincent et al.,[109] considering average values, that could be modified whether dispersion or transfer coefficients in various media become available from experimental data or from expertise of the company.

Table II-4. Transfer Coefficients for PEI calculation

	receiving medium		
physical state	***air***	***water***	***soil***
gas	0.95	0.05	0.001
liquid	0.5	0.35	0.002
solid	0.001	0.005	0.005
powder	0.1	0.85	0.005

Conclusively, PEI final score is calculated from Eq.10, in which maximum mass flow between inlet and outlet, m_i^{max}, specific for main product obtained, m_{prod}, multiplies the term related to PEI hazard class of chemical i, $H_{Cl_i}^{PEI}$, times the transfer coefficient for chemical i, d_i.

$$PEI = \sum_i \frac{m_i^{max}}{m_{prod}} \cdot 10^{H_{Cl_i}^{PEI}} \cdot d_i \qquad Eq.10$$

II.3.3.2 FRESHWATER TOXICITY POTENTIAL (FTP)

Since aquatic environment is often the final sink of many contaminants, due to direct immission or hydrologic/atmospheric processes, a particular concern has been conferred to the assessment of the chemicals impact on the aquatic habitat. Accordingly, the FTP has been introduced, since it is an indicator that inspects the environmental issues on freshwater sources due to polluting chemicals. This indicator, whose calculation is performed using Eq.11, considers the maximum mass of specific substance i, m_i^{max}, (that is inlet flow for raw materials, catalysts and solvents, outlet flow for products and byproducts) per mass of product, m_{prod}, times a chemical specific hazard class, $H_{Cl_i}^{FTP}$, related to in-silico estimated concentration acute toxicity threshold for aquatic organisms.

$$\text{FTP} = \sum_i \frac{m_i^{max}}{m_{prod}} H_{Cl_i}^{FTP} \qquad Eq.11$$

In this thesis, several methodologies have been used to evaluate the hazard class term $H_{Cl_i}^{FTP}$. One of them employed the web-based platform PreADMET[113] to predict the response of aquatic organisms, i.e. algae, daphnia magna, feathed minnow and medaka, even though no boundaries on the number of endpoints has been set, thus any number of aquatic organism can be implemented within this indicator. Experimental values of aquatic toxicity threshold (LC_{50}, EC_{50}, …) are preferred in comparison with in-silico estimated ones, however an abundance of compounds haven't been assessed yet in literature, due to the cost, long-time and ethical issues related to the animal experimental tests. The toxicity values generated by PreADMET have been inverted using Eq.12 in order to obtain their reciprocal, since a lower value of toxic concentration is indicative for more pollutant substances, but more toxic chemical should generate a greater impact on the overall PEI score: using the reciprocal of the concentrations preserves the ratio homogeneity among chemicals, while assigning a higher impact to more pollutant compounds.

$$H_{ij}^{FTP} = \frac{1}{C_{lim}} \qquad Eq.12$$

in which $H_{ij}^{FTP}$is the hazard score related to chemical i for aquatic organism j and C_{lim} is its concentration threshold value. It is essential that the entire set of chemicals involved adopts consistent units of measurement. For each aquatic organism, the hazard scores of the complete set of substances have been normalized from 0 to 1 using Eq.13, ensuring coherence among the variety of aquatic organism evaluated.

$$\widetilde{H}_{ij}^{FTP} = \frac{H_{ij}^{FTP} - H_{j,min}^{FTP}}{H_{j,max}^{FTP} - H_{j,min}^{FTP}} \qquad Eq.13$$

where $\widetilde{H}_{ij}^{FTP}$ is the normalized hazard score generated by chemical i for aquatic organism j and $H_{j,max}^{FTP}$ and $H_{j,min}^{FTP}$ represent the maximum and the minimum hazard scores of the entire set of chemicals for the specific aquatic organism j.

The substance i specific hazard class employed in Eq.11 is calculated using Eq.14, in which the normalized hazard scores for each aquatic organism j are summed up to attain $H_{Cl_i}^{FTP}$.

$$H_{Cl_i}^{FTP} = \sum_j \widetilde{H}_{ij}^{FTP} \qquad Eq.14$$

Beside PreADMET, another methodology employed for the calculation of $H_{Cl_i}^{FTP}$ involved the adoption of different techniques and software in order to obtain the evaluation of the impact on freshwater biota. First step contemplates the implementation of the molecular structure of chemical i into TURBOMOLE and, subsequently, in COSMOtherm in order to perform quantum mechanical

calculation and obtain the σ-surface of the molecule: this process will be deeply investigated in *Paragraph III.3.1*. The σ-profile is adopted by COSMOtherm to estimate several chemical properties which will be useful for the following steps of the methodology, i.e. vapor pressure, Henry's Law constant, solubility in water at 25°C, Kow, and Koc (soil sorption coefficient).[118] Bioconcentration Factor (BCF) is another chemical property estimated starting from Kow estimation results attained using COSMOtherm, as explained in the next *Paragraph II.3.3.3*. Since experimental values of lethal dose concentrations for a lot of different aquatic organisms are not usually available, an in-silico evaluation platform for LC_{50} has been considered for this purpose. In fact, the second step includes the adoption of ECOSAR,[119] which is an estimation platform based on SAR (Structure Activity Relationship) embedded within EPISuite software.[120] This platform employs structural information expressed by SMILES input to estimate the value of LC_{50} for various aquatic organisms belonging to different groups, including fish, algae, crustaceans and worms. The combination of the information obtained from the prediction of the first two steps are fundamentals for a proper estimation of the final freshwater impact, which is gained using USEtox.[77] The chemical properties predicted have been used as input for USEtox excel estimation of the specific chemical impacts on freshwater, which is the Ecotoxicity Effect Factor (EF_{eco}), measured as PAF (Potentially Affected Fraction of species) integrated over volume per unit mass of a chemical emitted ($PAF \cdot m^3 \cdot kg^{-1}$). The value of EF_{eco} has been assigned to $H_{Cl_i}^{FTP}$ and then the FTP score has been calculated using Eq.11.

II.3.3.3 BIOCONCENTRATION FACTOR (BCF)

An additional key parameter for the assessment of the effect of chemicals on the aquatic biota resides in the evaluation of the relationship between concentrations of chemicals in organisms against the one in the surrounding environment, which is mainly affected by the processes called bioconcentration and bioaccumulation. The former represents the bioconcentration capability of a chemical, defined as the ratio between its concentration in the organism exposed through non-dietary routes and the concentration in water at steady-state equilibrium under laboratory conditions,[121] while the latter refers to all the possible exposure routes (dietary, respiratory, dermal).[122]

The European REACH Regulation[123] identifies the risk assessment of bioaccumulative substances as a priority, thus there is an ongoing discussion about what is the most suitable surrogate parameter for bioaccumulation assessment. To date, BCF is still the reference endpoint under REACH for Persistent, Bioacculumative and Toxic (PBT) classification, albeit in the future, BCF may be substituted with bioaccumulation factor (BAF).[124] Alternatively, the octanol-water partition coefficient (Kow) can be used as a screening criterion,[122] as it infers where a compound is more likely to reside in a biphasic organic-aqueous equilibrium system.

Considering the methodology presented in this thesis, BCF can be employed by itself as an independent indicator or can be included within FTP, as shown in the case study of *Paragraph IV.2*.

Due to its central role, a plethora of in-silico QSAR (Quantitative structure-activity relationship) models have been developed in order to estimate BCF, refraining from the endorsement of animal testing.[121] In her study, Grisoni et al.[121] analyze nine QSAR models, aiming to define whether models based exclusively on Kow or more complex ones provide better BCF estimations. The conclusions underline how complicated model performed globally slightly better than simple models based on Kow only, however, since for the purpose of this analysis the simplicity of the evaluation is endorsed, the most reliable Kow-based model[123] provides BCF estimations that have been considered good enough for a great number of compounds.

The model adopted is part of the Technical Guidance Document on risk assessment[123] and suggests the following set of equations (Eq.15-18):

$$if\ logKow_i < 1 \rightarrow logBCF_i = 0.15 \qquad Eq.15$$

$$if\ 1 \leq logKow_i \leq 6 \rightarrow logBCF_i = 0.85 \cdot logKow_i - 0.7 \qquad Eq.16$$

$$if\ 6 < logKow_i < 10 \rightarrow logBCF_i = -0.2 \cdot (logKow_i)^2 + 2.74 \cdot logKow_i - 4.72 \qquad Eq.17$$

$$if logKow_i \geq 10 \rightarrow logBCF_i = 2.68 \qquad Eq.18$$

Since the Kow value is crucial for a good estimation of BCF, a deeper investigation on several in-silico techniques available for its quantification has been performed, aiming to identify the best existing practice regarding the evaluation of molecules with a molecular weight bigger than 150Da, that are rather common in pharmaceutical industry. This analysis will be presented in *Chapter III*.

II.3.3.4 PHOTOCHEMICAL OXIDATION POTENTIAL (PCOP)

PCOP or smog formation potential is one of the impact category adopted in the WAR algorithm proposed by Young.[98] It is determined by comparing the reaction rate (kOH) at which a unit mass of chemical i reacts with a hydroxyl radical OH to the rate at which a unit mass of ethylene reacts with OH. In this work, the normalization of the specific kOH using the ethylene reaction rate has been neglected, in favor of a direct utilization of the reaction rates times the mass flows of the chemicals involved in the process, as shown in Eq.19:

$$PCOP = \sum_i \frac{m_i^{max}}{m_{prod}} \cdot kOH_i \qquad Eq.19$$

where the maximum mass of specific substance i, m_i^{max}, per mass of product, m_{prod}, multiplies the chemical specific radical reactivity with OH, kOH_i. The reactivity with hydroxyl for each chemical has been predicted using MOPAC7 implementation within COSMOtherm.[125] This is based on a molecular orbital OH (MOOH) calculation (a semi-empirical AM1 calculations), which utilizes the lowest energy gas-phase conformation available.

II.4 Applying Data Envelopment Analysis (DEA) on indicators scores

The next step deals with the evaluation of the sustainability performance of all alternative routes involved in the case study considering the above selected indicators. In the realm of sustainability problems, economic, environmental, and social aspects need to be considered simultaneously. This characteristic brings the systems into multi-criteria decision-making problems, which are rather difficult to solve because inherent trade-offs usually arise between those criteria. Different approaches can be pursued in order to define the optimal solution among different designs adopting a multicriteria methodology.

Practitioners used to attribute, implicitly or explicitly, subjective weights to each specific impact for the sake of assigning different contributions, accordingly to relative importance given by stakeholders. However, this simple approach leads to relevant consequences which are worth to underline. A deep understanding of the trade-offs between different criteria needs to be achieved, otherwise a meaningful weights assignment expressing preferences might not depict what was intended to represent. Moreover, generation of weighting factors needs to be done by experts to limit as much as possible the introduction of biases in the methodology, which could be hard to remove further on and could lead to misleading results. Finally, due to the inadequate comprehension of the inherent trade-offs among indicators within the multicriteria problem itself, a guideline about how to improve suboptimal designs becomes challenging to provide.

In order to overcome these limitations, a different approach that allows to screen and select process designs has been chosen, which is Data Envelopment Analysis (DEA). DEA is a mathematical tool that employs linear programming (LP) techniques for the assessment and the evaluation of the efficiency of a group of homogeneous units considering multiple criteria simultaneously.[126] This technique, originally developed for the efficiency level evaluation of social systems, applies LP to quantify the relative efficiency of alternatives, adopting a non-parametric setting. Adoption of DEA in areas different than social one is well-established, from energy to environmental studies. Sustainable development analysis is a more recent application, in which life cycle assessment (LCA) principles are evaluated to assess the eco-efficiency performances of different technologies.[127–132]

Figure II-2 shows an illustrative example of how DEA works. Assume there are eight different process designs that need to be assessed in terms of cost and hazard toxicity potential. Given these designs (capital letters from A to H in the figure), the questions to answer is which of the eight are optimal and which are suboptimal. For the latter, the aim is also to establish improvement targets that if attained would make them optimal. Clearly, in this example, designs A, C, G and I are optimal (or efficient), as there is no other design that improves them simultaneously in both criteria. Conversely, the design D is suboptimal, as design C improves the former simultaneously in cost and hazard toxicity potential. By projecting the inefficient design onto the efficient frontier (e.g. from D to D'), it is possible to establish the values of the two objectives that such unit should attain so as to belong to the efficient frontier. Obviously, this type of analysis is almost straightforward when only two criteria are considered, but can become complex otherwise and more so as the number of indicators included in the analysis further increases. In DEA, linear optimization problems solver is employed independently for each unit, defined as Decision Making Unit (DMU), in order to define the set of most-efficient entities, which form the pareto-optimal frontier. Inefficient DMUs are further analyzed to identify the contributions that lowered the unit performance, providing individual targets to improve inefficient units and drive them closer to the efficient frontier.

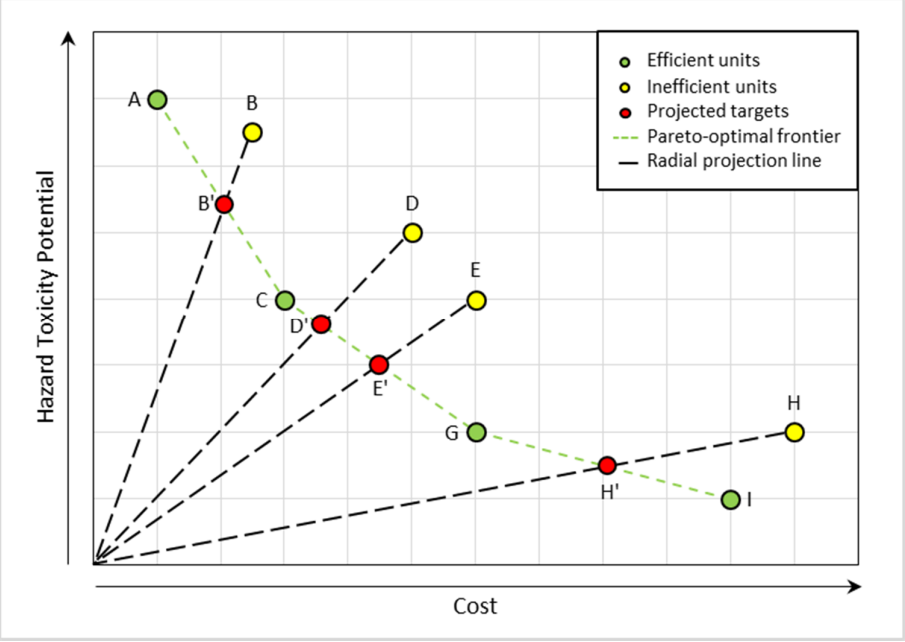

Figure II-2. Illustration of the DEA radial projection considering two inputs

Practitioners should bear in mind that in order to obtain reliable results using DEA, a reasonable balance between number of inputs (in this case number of indicators) and amount of DMUs (represented

by process designs) needs to be preserved. Cooper et al.[133] set the limit over which DEA is not able to provide reliable results, introducing the relationship of Eq.20.

$$DMUs \geq max\begin{Bmatrix} p \cdot s \\ 3 \cdot (p+s) \end{Bmatrix} \qquad Eq.20$$

where p and s are the number of inputs and outputs, respectively. Before performing DEA, experts should define the number of process alternatives to assess, identifying the number of indicators (inputs) to evaluate and the functional unit of production (outputs) accordingly to their needs. First, the maximum value between the two alternative formulas in Eq.20 brackets has to be identified, then it is necessary to verify that the number of DMUs is equal or greater the maximum value just determined.

II.4.1 DESCRIPTION OF DEA

DEA is a mathematical tool that exploit linear programming (LP) techniques to evaluate the relative performance of a group of systems (being process alternative routes in this work) considering several criteria simultaneously.

In the context of DEA, each system, called Decision Making Unit (DMU), is defined as a unit that consumes a certain quantity of resources (inputs) to produce an amount of outputs. The relative efficiency of any particular DMU_o is then calculated maximizing the ratio of the weighted sum of inputs to the weighted sum of outputs of that DMU_o through an optimization problem subjected to the constraints of the analogous ratio of the other DMUs being less than or equal to unity.[126] In essence, DEA identifies for each unit the best possible specific weighting factors attached to each input and output towards the maximization of the efficiency of each unit relatively to the others. For a given DMU_o, the original fractional problem of DEA is formulated to measure the relative efficiency of each DMU as follows:

Primal problem

$$\max_{u_r, v_i} \theta_o = \frac{\sum_{r=1}^{s} u_r y_{ro}}{\sum_{i=1}^{m} v_i x_{io}} \qquad M.1$$

$$subject\ to: \frac{\sum_{r=1}^{s} u_r y_{rj}}{\sum_{i=1}^{m} v_i x_{ij}} \leq 1; \qquad j = 1, \dots, n.$$

$$u_r, v_i \geq 0; \qquad r = 1, \dots, s; \qquad i = 1, \dots, m.$$

in which n is the total number of DMUs; o is the index for the DMU that is evaluated and takes integer values between 1 and n; j is the index for the other DMUs and varies in the range [1,n] ∈ ℕ; m is the number of inputs consumed by DMUj; x_{ij} represents the known amount of input i (i = 1,..., m) consumed by DMUj; s stands for the number of outputs generated by DMUj; y_{rj} is the given amount of output r (r = 1,..., s) generated by DMUj; v_i are the calculated linear weights assigned to input; u_r are the calculated linear weights assigned to output r; θ_o is the calculated relative efficiency score of DMU_o.

The above fractional problem features nonlinearity which can cause computational issues. The non-linear formulation of the problem can be translated into an equivalent linear programming (LP) model, enabling the adoption of more efficient solving techniques. Therefore, M.1 can be reformulated to ease its resolution as shown in the following model (M.2).

$$\max_{u_r, v_i} \theta_o = \sum_{r=1}^{s} u_r y_{ro} \qquad M.2$$

$$subject\ to: \sum_{i=1}^{m} v_i x_{io} = 1$$

$$\sum_{r=1}^{s} u_r y_{rj} - \sum_{i=1}^{m} v_i x_{ij} \leq 0; ; \quad j = 1, \dots, n.$$

$$u_r, v_i \geq 0; \qquad r = 1, \dots, s; \qquad i = 1, \dots, m$$

M.2 is solved for every single DMU providing each specific efficiency score (θ_o) as outcome of the model. Hence, if θ_o equals unity, the DMU will be considered efficient, otherwise, i.e. $\theta_o < 1$, that DMU will be deemed inefficient or suboptimal.

Dual problem

The dual problem is another model that is almost equivalent to the primal one but provides more insightful result interpretation, especially for the inefficient systems. In particular, it offers valuable guidance on the way to enhance the systems efficiency supplying efficient targets based on a reference set build on the efficient entities. The dual problem can be expressed as in M.3:

$$\min_{\lambda_j, S_i^-, S_r^+} \theta_o - \varepsilon \left(\sum_{r=1}^{s} S_r^+ + \sum_{i=1}^{m} S_i^- \right) \qquad M.3$$

$$subject\ to: \quad \sum_{j=1}^{n} \lambda_j x_{ij} + S_i^- = \theta_o x_{io}; \quad i = 1, \dots, m.$$

$$\sum_{j=1}^{n} \lambda_j y_{rj} - S_r^+ = y_{ro}; \qquad r = 1, \dots, s.$$

$$\lambda_j, S_i^-, S_r^+ \geq 0; \qquad \forall i, j, r; \qquad \theta_o\ unconstrained.$$

In which θ_o is the relative efficiency score of DMU_o, ε is non-Archimedean infinitesimal value to enforce the variables to be strictly positive, S_i^- are slack variables for input i or surplus amount of input needed to be reduced to become efficient, S_r^+ are slack variables for output r or additional amount of output to be increased to become efficient, λ_j are linear weights attached to every single DMU_j to form a linear combination.

This dual LP model offers valuable guideline to improve inefficient units through their projection to the Pareto frontier. The projected points on the Pareto-front (red dot points in Figure II-2) represent the efficient targets for improvements of the suboptimal units that would turn them optimal if achieved. The mathematical expression of each efficient target results from a linear combination of some selected efficient units, which assume a benchmarked peers function to the inefficient unit under investigation, i.e. $\sum_{j=1}^{n} \lambda_j\, x_{ij}$ for inputs and $\sum_{j=1}^{n} \lambda_j\, y_{rj}$ for outputs. Furthermore, the value of linear weights (λ_j) could reveal the extent to which such an inefficient unit attempts to imitate its reference peers. The improvement target calculated using the linear combination just mentioned are also adopted for the calculation of the improvement percentage that each DMU should achieve in order to become efficient, as shown in Eq.21:

$$imp_i = \frac{t_i - v_i}{v_i} \cdot 100 \qquad Eq.21$$

in which imp_i is the improvement percentage of DMU i, t_i represent the efficient target of DMU i in the specific input (in this case indicator) and v_i is the original value of DMU i for the same input.

The calculation of such efficient targets, demands the prior establishment of two main components, i.e. the envelopment of the Pareto front and the projection from the suboptimal units onto

the front. Two different approaches have been developed in order to construct the Pareto front, each one assuming a peculiar dependence upon the assumption of returns to scale. Models based on M.1 to M.3 assume constant returns to scale (CRS), meaning that changes in inputs are proportionally transferred to variations in outputs. These are known as CCR model that is the original DEA model proposed by Charnes, Cooper, and Rhodes.[126] Another well-established DEA model is the BCC model (Banker-Charnes-Cooper)[134] that aims to extend CCR model introducing a variable returns to scale (VRS), meaning that a variation in inputs does not necessarily lead to a proportional change in outputs. The complexity of the VRS model is grown in comparison with CRS one, as the construction of the VRS model requires an additional convexity constraint ($\sum_{j=1}^{n} \lambda_j = 1$) that explicitly dictates the piecewise linear and concave characteristics of the VRS frontier.

Moreover, higher efficiency score can be achieved by improving following the projecting trajectory towards the Pareto front. In the context of DEA, this improvement can be carried out pursuing different approaches, focusing on the input or on the output minimization. The former considers radial projection with inputs reduction, while at least the same amount of outputs is still preserved. Notwithstanding, the latter aims to minimize outputs while maintaining the amount of inputs almost at a certain level, called output-oriented model. Since the majority of interested criteria for sustainability evaluation have been categorized as inputs, input-oriented model is the one employed in this study.

The aforementioned models of DEA provide guidance to practitioners concerning the identification of efficient DMUs and the calculation of efficiency targets for inefficient units, nevertheless it is inadequate to fulfill a task that practitioners still concerns. Since all the efficiency scores for units belonging to the Pareto front in the traditional DEA are calculated as 1, it is rather difficult to rank them. As soon as they have identified the efficient process designs, how can they discriminate which is the one that performs best among the optimal DMUs?

In order to satisfy this demand, another extension model, that could potentially improve the discriminatory power of DEA, has been implemented, called super-efficiency model. The model contains essentially the same equations as in M.3, except that now both constraints exclude the efficient unit being assessed from the summation terms on the left hand side.[133,135] Let j' be the efficient unit for which the super-efficiency is calculated, then the constraints in model M.3 would be changed as follows:

$$\sum_{j=1, j\neq j'}^{n} \lambda_j x_{ij} + S_i^- = \theta_{j'} x_{ij'}; \quad i = 1, \dots, m. \qquad M.4$$

$$\sum_{j=1, j\neq j'}^{n} \lambda_j y_{rj} - S_r^+ = y_{rj'}; \quad r = 1, \dots, s.$$

The super-efficiency model quantifies the extent to which the efficient unit being assessed over-performs the Pareto front that is constituted from the rest of the efficient units. It provides an indication on which efficient unit, if removed from the overall analysis, would worsen most the envelopment of the Pareto front. Hence, after the application of this model to every efficient unit, it provides a super-efficiency score $\theta_{j'}$ that is always greater than or equal to one, which can be used to further discriminate among efficient units, identifying the most efficient unit as the one with $\theta_{j'}$ maximum value.

II.5 IMPROVING SUB-OPTIMAL DESIGNS

It is now essential to understand the behavior of the systems under investigation, analyzing the factors that led indicator scores towards higher values, meaning a higher impact and, therefore, a suboptimal design. This retrofit analysis examines the intrinsic calculation of the indicators chosen, underlining the major contributions to the final values obtained. The results gained through DEA provide a first hint on the identification of the indicators which contributed most to the suboptimal condition, calculating the percentage of improvement for each indicator in order to become as efficient as the optimal one. Each percentage represents the enhancement that each DMU should reach in order to move its position to the Pareto front. Aiming to increase the performances of DMUs, it is necessary to further evaluate the influence of the variables that reside in the calculation of indicators, highlighting which ones will be worth to modify to obtain a relevant improvement on the final score. The results of this sensitivity analysis depends on the indicators chosen, therefore it is not possible to state a priori the most affecting decision variables for every case study. In this thesis, some variables have been fixed based on the experimental data found in the patents under investigation, *e.g.* purity of streams, ratio between amounts of chemicals involved, and yield of reactions. Moreover some other variables were fixed among the DMUs for the sake of evaluating the influence of specific contributions, which were more significant from an industrial perspective.

In the next chapter the reader will find a description of several molecular simulation techniques used for the estimation of a specific parameter, i.e. Kow, adopted for the calculation of an environmental sustainability indicator.

CHAPTER III

ESTIMATION OF KOW USING MOLECULAR MODELING TECHNIQUES

III.1 OCTANOL WATER PARTITION COEFFICIENT (KOW)

DEFINITION

A pure substance dissolved in a biphasic solution of two partially miscible solvents in equilibrium distributes itself between them in accordance with the mutual relative interaction properties. The ratio of solute concentrations in the two phases describing this behavior is known as partition coefficient.[136] Partition coefficient of a substance between an organic phase in intimate contact with an aqueous one is a well-established measurement of the hydrophilicity of a chemical.

Partition coefficients have been adopted to forecast the distribution of drugs within the body (since hydrophobic substances are more likely to reside within hydrophobic lipid bilayers of cells, while hydrophilic drugs mainly distributes in aqueous blood serum),[137] as well as to mimic biologic activity of pollutant chemical, as they are able to estimate the tendency to pass through cells membrane.[138] Aiming to select the most appropriate set of solvents to estimate the partition coefficient of such systems, the couple water/1-octanol arise as the most trustworthy.[138] The amphiphilic character typical of lipids in biological membranes is a peculiarity of 1-octanol as well,[138] provided by its two main areas of constitution: the polar head and the neutral tail (Figure III-1). Indeed, 1-octanol results primarily lipophilic due to the length of its alkyl chain, while its hydrophilic hydroxyl group provides for delimited charged areas in which hydrogen bond with water molecules can occur.

Figure III-1. Representation of 1-octanol amphiphilic behavior using σ-surface in which blue represents the positive area, red the negative one and green the neutral region

For the time being, an ecological risk assessment for organic compounds without consideration of Kow value seems impossible.[139] In natural systems, chemical compounds reach a specific target, e.g. a cell, from an external dilute solution following a random path. This step is followed by the transfer of the substances inside the organism through a relatively slow process of diffusion or permeation via cell membranes, which represents the bottleneck of the entire process and is controlled by the partitioning

trend of the compounds between a polar aqueous phase and a neutral organic one. The molecular structures of the compounds are affecting the overall process, since, as soon as one compound moves into the internal of the cell, it goes through a series of chemical reactions eliciting a biological response.[140] Hence, the central role of octanol-water system in environmental matter is due to its peculiarity of being a reliable indicator for the quantification of the rate controlling step during the interaction of a compound in a biological system, which is the transfer from an aqueous-like phase to organic-like one inside organisms.[140]

Kow is a dimensionless quantity, whose numerical expression is shown in Eq.22, in which, for a given temperature, $[X]_s^o$ represents the concentration of the solute in the octanol phase and $[X]_s^w$ is the concentration of the same compound in the aquatic one.

$$Kow = \frac{[X]_s^o}{[X]_s^w} \qquad Eq.22$$

Furthermore, the chemical potential of the solute in water and in octanol can be expressed by Eq.23 and Eq.24, respectively.

$$\mu_s^w = \mu_s^* + RT \ln \gamma_s^w \phi_s^w \qquad Eq.23$$

$$\mu_s^o = \mu_s^* + RT \ln \gamma_s^o \phi_s^o \qquad Eq.24$$

where μ_s^* represents the chemical potential of the pure liquid solute, R is the ideal gas constant, T is the temperature of the system expressed in Kelvin, γ_s^w and γ_s^o are the solute activity coefficients on a volume fraction basis ($\gamma_s \to 1$ as $\phi_s \to 1$) in water and octanol, while ϕ_s^w and ϕ_s^o are respectively the volume fractions of solutes in water and octanol.[141]

Since the equilibrium condition requires that $\mu_s^w = \mu_s^o$ is reached, Eq.23 and Eq.24 can be condensed in Eq.25:

$$\gamma_s^w \phi_s^w = \gamma_s^o \phi_s^o \qquad Eq.25$$

Being $\phi_s = [X]_s V_s$, in which V_s is the partial molar volume of the solute, the combination of Eq.22 and Eq.25, gives rise of the following Eq.26.

$$\log Kow = \log \gamma_s^w - \log \gamma_s^o \qquad Eq.26$$

Assuming $V_s^w \approx V_s^o \approx V_s^*$, with the last term representing the molar volume of pure liquid solute, Eq.26 highlight how Kow can be expressed as a free energy function, hence its intrinsic connection with the energetics transfer between two phases.[140]

A study about the prediction of Kow based on experimental data (Pfizer database)[142] report values for small organic molecules from 10^{-3} to 10^{9}, exhibiting a range of twelve order of magnitude, which is rather inconvenient. Thus the logarithmic expression of Kow emerged as the best representation of this property, since it varies in a range from -3 to 9 for small organic molecules. It is also relevant to point out that the calculation of logKow from the ratio of solubilities in octanol and water considered separately is rather inaccurate, as the reported in Sijm et al. work,[143] since the mutual solubility of water and 1-octanol affects the distribution of solutes. While the solubility of octanol in water is very low, the equilibrium solubility of water in octanol at room temperature is reported to be about 5% in mass.[144]

Since Kow represent the distribution of nonpolar organic compounds between water and natural solids (e.g., soils, sediments and suspended particles) or living organisms, the values of logKow<1 will be assumed by substances that are mainly hydrophilic, so they tend to exhibit higher water solubility, smaller soil/sediment adsorption coefficients, and smaller bioaccumulation factors for aquatic life. Vice versa, chemicals with logKow>4 result mainly hydrophobic, revealing an opposite behavior.[145,146]

Different experimental methodologies have been proposed to measure Kow value,[147–151] until a standardization has been reached by an OECD protocol[152] followed by a validation via ring test.[153]

The number of chemicals available in the market is progressively increasing: in the Chemical Abstract Service Register[154] only, more than 133 million substances are registered up to date (October 2017), and new ones are included on a daily basis. The numerous chemical properties associated to each substance require an abundance of experimental tests to be performed for their total assessment, leading to an impractical series of experimental procedures. Therefore, beside the experimental procedures which are rather time-demanding and need a specific experimental set-up and expertise, in-silico methodologies for calculation of molecular properties, including Kow, arise as a valuable alternative.

III.2 In-silico methodologies for Kow estimation

Since octanol-water partition coefficient has always played a crucial role in the estimation of a wide range of equilibrium, repartition and toxicological properties of a compound, the development of in-silico methodologies has been thriving in the last two decades.

Some of the different Kow estimation techniques, which are based on distinct fundamentals, will be presented in the following paragraphs.

III.2.1 FREE SOLVATION ENERGY RELATED METHODS

Since partition coefficients are thermodynamic properties related to Gibbs free energies of solvation in two different phases at a fixed temperature T, it is possible to calculate the one related to the water-octanol system adopting the following expression (Eq.27):

$$logKow = \frac{\Delta G^{w}_{solv} - \Delta G^{o}_{solv}}{2.303RT} \qquad Eq.27$$

where ΔG^{w}_{solv} is Gibbs free energy of solvation in water, and ΔG^{o}_{solv} is the Gibbs free energy of solvation in water-saturated octanol phase. Thus, the octanol-water partition coefficient calculation requires an estimation of the free energies of solvation in both aqueous and organic phases.

Various methodologies have been adopted, coupled with Monte Carlo or Molecular Dynamics (MD) techniques, to obtain reliable free solvation energy computations,[155] which are briefly listed and described underneath:

- *Thermodynamic integration (TI)*: since a derivative of the free energy can be determined directly as an ensemble average from a simulation, this method samples this derivative along a path between two reference states, allowing the free energy to be determined by numerical quadrature.[155] Indeed, being the free energy a function of the Boltzmann-weighted integral over phase space coordinates of the system, the free energy difference between two states cannot be calculated directly, but rather by defining a thermodynamic path between the reference states and integrating over ensemble-averaged enthalpy changes along the path. Considering two systems, A and B, with potential energies U_A and U_B, the potential energy in either system can be calculated as an ensemble average over configurations sampled from a MD or Monte Carlo simulation with proper Boltzmann weighting,[156] using the new potential energy function defined as:

$$U(\lambda) = U_A + \lambda(U_B - U_A) \qquad Eq.28$$

where λ is defined as a coupling parameter with a value between 0 and 1, and thus the potential energy as a function of λ varies from the energy of system A for $\lambda = 0$ and system B for $\lambda = 1$.

- *Free Energy Perturbation (FEP) method:* coupled with MD, FEP has been adopted by Best et al.[157] in order to obtain insights into the structural peculiarities and dynamic behavior of a set of small organic solutes dissolved in water and water-saturated 1-octanol. The FEP theory, introduced by Zwanzig,[158] is based on statistical mechanics and is used in computational chemistry for computing free energy differences using molecular dynamics or Monte Carlo simulations. Considering the free energy cycle of Figure III-2, in order to obtain the Kow value of solute A, initially the free energies of solvation are determined in the aqueous phase, $\Delta G^{w}_{solv}A$, and in the water-saturated octanol phase, $\Delta G^{o}_{solv}A$. The difference between the solvation free energies is the free energy of transfer, $\Delta G_{Tr}A$, representing the energy related to the transfer of solute A from the different solvents. This process is illustrated for two different solutes, A and B, by the vertical paths 1 and 3 in Figure III-2.

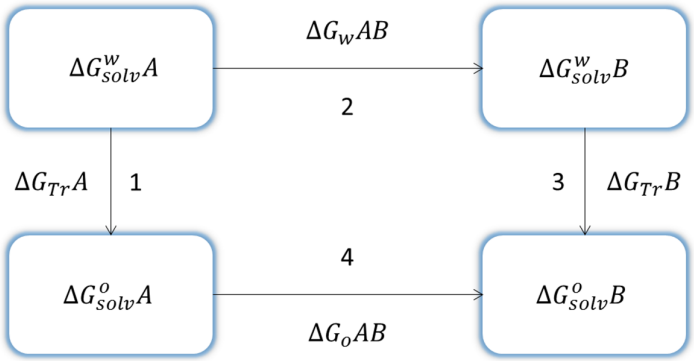

Figure III-2. Thermodynamic free energy cycle for the calculation of octanol-water partition coefficient.

The free energy of transfer for solutes A and B is related to octanol-water partition coefficients for A ($logKow(A)$) and B ($logKow(B)$) through Eq.29 and Eq.30, respectively, where R is the gas constant and T is the temperature.

$$\Delta G_{Tr}A = -2.3RTlogKow(A) \qquad Eq.29$$

$$\Delta G_{Tr}B = -2.3RTlogKow(B) \qquad Eq.30$$

Since it is difficult to compute absolute free energies of solvation using FEP simulations,[159] scientists followed a different route for Kow calculation, taking advantage of the free energy peculiarity of being a state function. Therefore, since the calculation of the relative free energy of solvation in which solute A is slowly mutated into solute B ($\Delta G(AB)$) is more computationally feasible, the relationship expressed by Eq.31 can be exploited:

$$\Delta G_o AB - \Delta G_w AB = \Delta G_{Tr} B - \Delta G_{Tr} A = \Delta\Delta G_{Tr} AB \qquad Eq.31$$

The relative free energies of solvation of A and B can be used to calculate a relative free energy of transfer $\Delta\Delta G_{Tr} AB$. This allows the relative partition coefficient ($\Delta logKow$) for solutes A and B to be calculated from the direct relationship illustrated in Eq.32.

$$\Delta\Delta G_{Tr} AB = -2.3RT\Delta logKow \qquad Eq.32$$

- *Umbrella sampling:* this method employs biasing potentials to bias the simulated system into regions of configurational space that otherwise are very sparsely sampled.[160] This approach allows to cover the two reference systems in a single simulation run and to compute the free energy difference between them. The choice of biasing potentials is, however, not known a priori; often a series of partially overlapping harmonical potentials is used. The results from one simulation per biasing potential in the series are pasted together based on areas with relatively poor statistics, therefore yielding uncertain results in finite length simulations.[161] Adaptive umbrella sampling is an improvement of umbrella sampling, which implies a series of simulations where the biasing potential is updated to approximate the potential of mean force over a subset of coordinates. This allows uniform sampling of the coordinates of interest and therefore also enhanced efficiency and accuracy in the determination of free energy differences.[155]
- *Expanded Ensemble method:*[162] is similar to umbrella sampling, as it introduces a sampling system procedure in which the two reference states and the region between them are mapped. It is achieved by considering the coupling parameter (λ in Eq.28) as an extra coordinate and sampling over this coordinate. Additionally, a biasing potential along this coordinate is introduced, using a series of trial runs to calculate its balancing factors in a procedure analogous to adaptive umbrella sampling. This method, originally developed within Monte Carlo methodology and later adapted to MD simulation techniques, requires only one single run (after the short trial runs), while most other methods need a series of repeated computer simulation runs to obtain the free energy value.[162]

III.2.2 GIBBS ENSEMBLE MONTE CARLO (GEMC) METHOD

The Gibbs Ensemble Monte Carlo (GEMC) method[163] is ideally suited for the evaluation of partition coefficient, as it allows a setup analogous to the experimental situation. GEMC utilizes two (or more) separate simulation boxes which are in thermodynamic contact, whilst there is not an explicit interface, as shown in Figure III-3.

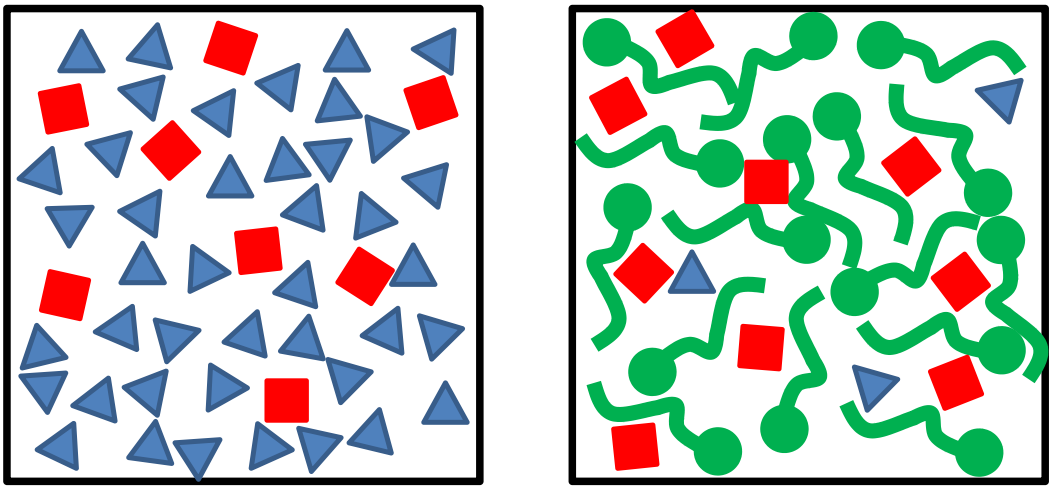

Figure III-3. Schematic representation of the Gibbs ensemble setup for the investigation of the partitioning of one solute (red squares) between a water-saturated 1-octanol phase (green circles with tail segments) and a water phase (blue triangles)

As a result, for a given state point the properties of the coexisting phases, such as the mutual solubilities of the two solvents and the partitioning of solute molecules, can be determined directly from a single simulation. The molecules are subjected to six different types of Monte Carlo moves to sample phase space efficiently: translational, rotational, Configurational-bias Monte Carlo (CBMC)[162] conformational, CBMC swap, CBMC switch, and volume moves. Monte Carlo methods are a class of computational algorithms based on repeated random sampling to identify the most probable event or the optimal condition depending on numerous variables simultaneously, overcoming the issue related to computational calculation limits. The fundamental idea is using randomness to solve problems that might be deterministic in principle, imitating Brownian motions and molecular behavior. The first three types of moves involve only a single molecule in a given box and ensure thermal equilibration. The CBMC swap move involves the particle exchange of a (solute or solvent) molecule from one phase to the other, thereby equalizing the chemical potentials of each species in the two phases. If the analysis includes more than one solute, during a CBMC switch move, molecule A is regrown as molecule B in one box, and B is regrown as A in the other box. This move has a much higher acceptance rate than the straightforward CBMC swap move of the larger B molecule and is used to equalize the differences in chemical potentials of A and B in

the two boxes. Finally, volume moves involving an external pressure bath lead to mechanical equilibrium between the two phases.[163] The main benefits provided by CBMC/GEMC simulations over the previous methods are multiple:

- in both the experiment and the CBMC/GEMC simulations the Gibbs free energy of transfer is directly determined from the ratio of solute number densities in the two phases,[164] while TI and FEP calculations estimate the difference in excess chemical potentials related to a specific standard state;
- the number density ratio is a mechanical property that can be determined very precisely from CBMC/GEMC simulations leading to small statistical errors in ΔG;
- the composition of the two solvent phases does not need to be specified in advance in CBMC/GEMC simulations, because the distribution of solvent molecules is also sampled via swap moves. Whereas TI and FEP calculations can only be carried out at a fixed composition (e.g., using the experimental data to construct a water-saturated 1-octanol phase), that might not correspond to a proper thermodynamic state for the force field used in the calculations.[163]

III.2.3 QSAR/QSPR METHODOLOGIES

Quantitative Structure-Activity Relationship (QSAR) and Quantitative Structure Property Relationship (QSPR) allow the prediction of Kow (and numerous other properties) of a given compound as a function of its molecular structure. Essentially, new and untested chemicals characterized by similar molecular features as substances adopted in the development of QSAR/QSPR models are likewise assumed to exhibit analogous activities/properties. The development of QSAR/QSPR model typically constitutes of two major steps, i.e. the identification of molecular structure and influencing descriptors and the correlation of molecular descriptors with observed activities/properties using various techniques.[165] Concerning the former, molecular descriptors are the final result of a logic and mathematical procedure which transforms chemical information encoded within a symbolic representation of a molecule into a useful number.[166]

The definition of the appropriate molecular descriptors is crucial for a reliable estimation of the desired property, thus a molecular structure needs to be divided into suitable subgroups, following the so called “substructure approaches.” Molecules are cut into atoms (atom contribution methods) or groups (fragmental methods) summing the single-atom or fragmental contributions in order to obtain the final Kow prediction. However, molecules are not mere collections of fragments or atoms, therefore fragmental methods apply correction rules coupled with molecular connectivity, as shown in Eq.33:

$$Kow = \sum_{i=1}^{n} a_i \cdot f_i + \sum_{j=1}^{m} b_j \cdot F_j \qquad Eq.33$$

where a_i is the number of fragment of type i, f_i is the fragmental quantification constant, F_j is the correction factor and b_j stands for the frequency of F_j.

On the contrary, most atom contribution methods work without correction factors, as reported in Eq.34:

$$Kow = \sum_{i=1}^{n} n_i \cdot a_i \qquad Eq.34$$

in which n_i is the number of atoms of element i and a_i is the constant related to the contribution of element i to the final property.

Fragmentation of a molecule can be somewhat arbitrary and any fragmentation approach has its own benefit and drawbacks. Fragments larger than a single atom can be defined, comprehending significant electronic interactions within one fragment, which represents a main advantage of using fragments, particularly for big molecules. An advantage of atom contribution methods resides in the avoidance of ambiguities, although a huge number of atom types is needed to describe a reasonable set of molecule.

The molecular descriptors mentioned so far are strictly related to the physical structure of the molecule, however, a different group of molecular descriptors that concern the electronical and quantum-mechanical properties of molecules, have been adopted in the QSAR field. New *ab initio* and semi-empirical methods supply reliable quantum-chemical molecular descriptors in a relatively short computational time frame, therefore quantum chemical methods can be applied to QSAR by direct derivation of electronic descriptors from the molecular wave function, which guarantees a more accurate and detailed description of electronic effects compared to empirical methods based on standard molecular descriptors.[167] The quantum-chemical descriptors employed in QSAR methodologies, listed in a comprehensive review by Karelson et al.,[167] have been gathered into several macrogroups regarding the main electronic properties they aim to quantify: Atomic Charges, Molecular Orbital Energies, Frontier Orbital Densities, Superdelocalizabilities, Atom-Atom Polarizabilities, Molecular Polarizability, Dipole Moment and Polarity Indices, Energy, and Others. Several papers adopting quantum-chemical molecular descriptors for Kow estimation have been published, using both structure-related and quantum-mechanics approaches[168,169] or quantum-mechanical ones exclusively.[170–173]

Once the contribution of each molecular descriptor has been defined, the QSAR model needs to quantitatively discern the relationships between the independent variables (e.g. molecular descriptors)

and the dependent variables (e.g. biological/chemical properties of interest), which is called Multivariate Analysis. Nowadays, several techniques are adopted to build the model that describe best this relationship:

- *Multiple Linear Regression (MLR):* the classical approach is a linear regression technique typically involving the establishment of a linear mathematical polynomial equation relating the variation of biological/chemical properties as a function of the variations of the molecular substituents present in the molecular data set, as shown in Eq.35:

$$y = a_0 + a_1x_1 + a_2x_2 + \ldots + a_nx_n \qquad Eq.35$$

  in which y is the dependent variable (e.g. biological/chemical property under study), a_0 is the baseline value for the chemical data set, $a_1, \ldots, a_n$ are the regression coefficients calculated from a set of training data (which is a fraction of the total data available for the development of the model) in a supervised manner where the independent and dependent variables are known. Such linear approach is suitable for systems in which the phenomenon of interest exhibits a linear relationships with molecular descriptors, as reported in external works.[174–177] However, since sometimes this constrain is not satisfied and the relationships between dependent/independent variables may be nonlinear in nature, a call upon the use of non-linear approaches in order to properly model such properties becomes essential.

- *Artificial neural network (ANN)* is a pattern recognition technique that aims to resemble the inner workings principles of the brain which is essentially composed of interconnected neurons exchanging information. Such structure is emulated by ANN's architectural design where neuronal units are interconnected to one another using various schemes, among which a well-established one is based on a three-layer feedforward network (input layer, hidden layer and output layer). The role of neurons belonging to input layer is to acquire the information of the independent variables and bring them into the ANN system: therefore the number of neuronal units available in the input layer needs to match to the number of independent variables in the data set. The hidden layer processes the information assigning different numerical values known as weights (which resemble the connection between neurons), recognizing a pattern and providing system behavior predictions that are passed to the output layer. In a backpropagation algorithm, the error calculated from the difference between the

predicted value and the actual value is obtained in order to define if its value is lower than a previously set tolerance threshold. If this is the case, then the learning process will stop, otherwise signals will be sent backwards to the hidden layer for further processing and weight readjustments. This is performed iteratively until a solution is reached and learning is terminated.[165] The reader can find some remarkable examples of this technique applied on the evaluation of Kow consulting some excellent resources elsewhere.[178–181]

- *Partial Least Square Regression (PLSR)* is a well-established regression methodology and a profusion of articles have been adopted this technique in the last two decades.[182–184] Unlike MLR, PLSR can handle data strongly collinear (correlated), noisy, with numerous independent variables (i.e. molecular descriptors), and characterized by simultaneously model several response variables (i.e. dependent biological/chemical properties). Each model parameter is iteratively estimated as the slope of a simple bivariate regression (least squares) between a matrix column (or row) as the x-variable, and the other parameter vector as the y-variable. So, for instance, the PLS weights, w, are iteratively re-estimated using Eq.36:

$$w = \frac{X'u}{u'u} \qquad Eq.36$$

being X the matrix of independent variables and u a vector of variables.

The "partial" in PLS indicates that this is a partial regression, since the x-vector (u above) is considered as fixed in the estimation. This also shows that we can see any matrix–vector multiplication as equivalent to a set of simple bivariate regressions. This provides an intriguing connection between two central operations in matrix algebra and statistics, as well as giving a simple way to deal with missing data.[185]

- *Support Vector (SV) machines* are supervised learning models with associated learning algorithms that analyze data in order to perform a regression analysis, as reported in remarkable manuscripts.[186–188] The regression path follows the steps shown in Figure III-4 which has been adapted from Smola et al.[189].

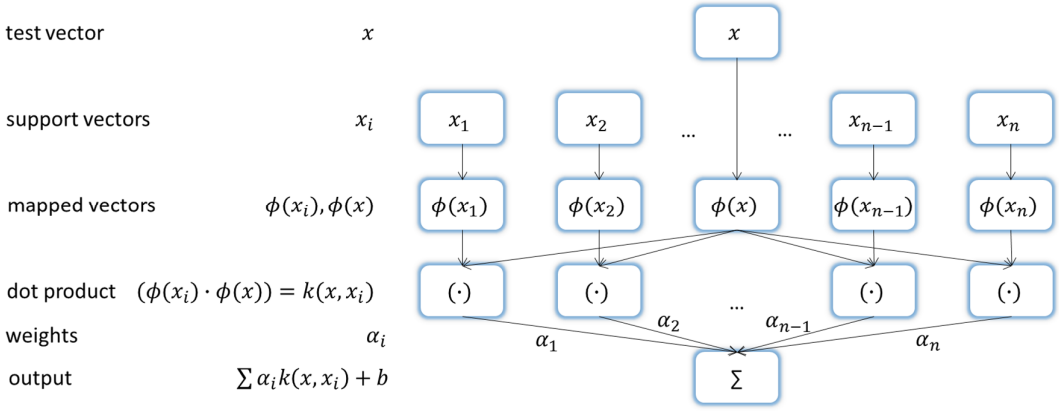

Figure III-4. Structure of a regression machine based on support vector algorithm

The input pattern (i.e. a new molecule for which a prediction need to be performed) is mapped into feature space (already constructed using known chemicals) by a map ϕ. Then dot products are calculated with the images of the training patterns under the map ϕ and added up using the weights α_i: this, plus the constant term b yields the final prediction output. The overall process described here is very similar to regression in a neural network, although in the SV the weights α_i are a subset of the training patterns.[189] In SV machine regression, the goal is to find a function whose deviation from the actually obtained targets y_i for all the training data is below a value ε through the solution of an optimization problem, obtaining a model as flat as possible. The errors are negligible as long as they are less than the tolerance ε, meaning that any deviation larger than ε requires the generation a new solution for the minimization problem. The influence of ε on the shape of the regression function is shown in Figure III-5, which has been adapted from Smola et al.[189]

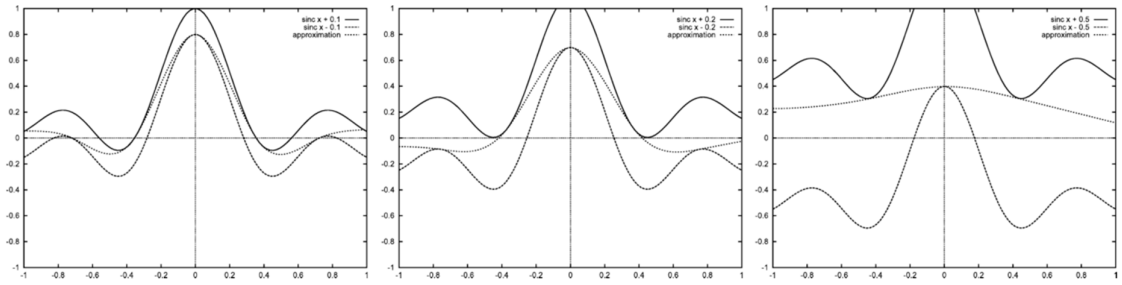

Figure III-5. Approximation of the function sinc x with precisions ε= 0.1, 0.2, and 0.5 from left to right, respectively. The solid top and the bottom lines indicate the size of the ε-tube, the dotted line in between is the regression.

- *Gene Expression Programming (GEP)* was invented by Ferreira[190] and was developed from genetic programming, which is a technique whereby computer programs are encoded as a set of genes that are then modified using an evolutionary algorithm (usually a genetic algorithm, GA). GEP mainly includes two sides, i.e. the chromosomes (which represent the molecular descriptors) and the expression trees, ETs (which symbolize the connection of molecular descriptors with molecular properties). The process of information of gene code and translation is very simple, such as a one-to-one relationship between the symbols of the chromosome and the functions or terminals they represent. The rules assigned to GEP determine the spatial organization of the functions and terminals in the ETs and the type of interaction between sub-ETs.[191] The purpose of symbolic regression or function finding is finding an expression able to provide a good explanation of the relationship between independent and dependent variable. The GEP procedure comprehends various steps, among which the first one is to choose the fitness function f_i, expressed by Eq.37:

$$f_i = \sum_{j=1}^{n} \left(R - \left| \frac{P_{(ij)} - T_j}{T_j} \cdot 100 \right| \right) \qquad Eq.37$$

in which R is the selection range, $P_{(ij)}$ is the value predicted by the individual program i for fitness case j (out of n fitness cases), and T_j is the target value for fitness case j. The absolute value term, corresponding to the relative error, is called the precision and if the error is smaller than or equal to the precision then the error becomes zero. Thus, for a good match between the model and the real behavior, the absolute value term is zero and $f_i = f_{max} = nR$. The second step consists of choosing the set of terminals T and the set of functions F to create the chromosomes. For instance, the functions F are built on the molecular descriptors, while the terminal set are represented by the property prediction. The third step is to choose the chromosomal architecture, i.e. the length of the head and the number of genes. The choice of the linking function represents the fourth step of the procedure. The last major step is to choose the set of genetic operators that cause variation and their rates, causing the "evolution" of the system by applying (among all possible) crossover, mutation, or rotation. These processes are repeated for a pre-specified number of generations until a solution is obtained. In the GEP, the fitness of each individual is the main factor for the selection of the entities kept for the next iterative generation, as determined by roulette-wheel sampling with elitism, which guarantees the screening of the best individual and their replication to the next generation. The results achieved by several studies[192–194] are satisfactory and show a promising use in the nonlinear structure-activity/property relationship correlation area, despite GEP is

congenitally defective due to some concerns about the reproducibility of the predicted values and regarding the trend of deducing very complex equations..[191]

- *Project Pursuit Regression (PPR)* was developed by Friedman and Stuetzle[195] as a powerful tool for seeking the interesting projections from high-dimensional data into lower dimensional space by means of linear projections. PPR approach overcomes the issues related to dimensionality because it relies on estimation in at most trivariate settings, avoiding the computational limitations of other existing nonparametric regression procedures. Moreover, interactions of predictor variables are directly considered, since general smooth functions are adopted in order to model the linear combinations of the predictors. A peculiarity of PPR is that the results of each interaction can be represented graphically and the graphical output can be used to modify the major parameters of the procedure, i.e. the average smoother bandwidth and the terminal threshold. PPR combines both ideas of projection of independent variables into the dependent variables space and the pursuit of the finite sequence of projections that can reveal the most interesting structures of the data. In a typical regression problem, PPR aims to approximate the regression pursuit function $f^{(p)}(x)$ by a finite sum of ridge functions with suitable choices of $\alpha_i$and f_i (Eq.38):

$$f^{(p)}(x) = \sum_{i=1}^{p} f_i(\alpha_i^T x) \qquad Eq.38$$

in which α_i values are m × n orthonormal matrices representing the coordinates of the n projected data onto the m-dimensional space of dependent variables and p is the number of ridge functions. In various QSAR/QSPR studies,[196–198] PPR was employed as a regression method for different chemical/biological properties, indicating that PPR is a promising regression method in QSAR/QSPR studies, especially when the correlation between descriptors and activities or properties is nonlinear.[191]

- *Local Lazy Regression (LLR):* the common trend of QSAR/QSPR models is capturing the global structure-activity/property relationships considering the whole dataset. In several cases, there may be specific groups of molecules which exhibit some peculiar features related to their activity or property, exhibiting what is so called a local structure activity/property relationship. Traditional QSAR models may not recognize such local relationships, hence LLR has been recognized as an excellent approach, enabling the practitioner to extract a prediction by locally interpolating the neighboring examples of the query which are deemed relevant considering to a distance measure, rather than the whole dataset. Therefore, the basic core of LLR resides on the assumption that similar

compounds have similar activities, linking any modification in the chemical structure to a simultaneous change on the activity of the compound. For one or more query points, LLR estimates the value of an unknown multivariate function on the basis of a set of possibly noisy samples of the function itself. Each sample is an input/output couple composed by an input vector and an output relative quantification value. For each query point, the estimation of the input is obtained by combining different local models comprehending a combination of polynomials of zeroth, first, and second order that fit a set of samples in the neighborhood of the query point, accordingly to either the "Manhattan" or the "Euclidean" distance. Furthermore, it is also possible to assign weights to the different directions of the input domain in order to quantify their importance in the computation of the distance. The number of neighbors used for identifying local models (identified using the recursive least-squares algorithm) is automatically adjusted on a query-by-query basis through a leave-one-out cross-validation of models, each fitting a different number of neighbors. The model assumes a linear dependence between the dependent variable and the predictor variables in a small region around x. Then LRR determines the points around $X_{NN(x)}$ and builds a regression model with only the points in $NN_{(x)}$using the least-squares method and minimizes the squared residuals for the region using this model. This technique prevents the construction of multiple models for each point in the training set beforehand, since for every a query point, the approach builds a representative predictive model. Hence, LLR equation is described below (Eq.39):

$$\beta_x = (X_{NN(X)}^T X_{NN(x)})^{-1} X_{NN(X)}^T Y_{NN(x)} \qquad Eq.39$$

in which $X_{NN(x)}$ is the matrix of independent variables (molecular descriptors), $Y_{NN(x)}$ represents the column vectors of the dependent variables (chemical properties) for the molecules in the neighborhood of the query point and β_x is the column vectors of regression coefficients. Since LLR does not require the development of any a priori model, is suitable for large data sets, where using all of the observations can normally be time-consuming and even lead to overfitting. Meanwhile, the construction of a regression model for each query point, enables to extract meaningful structure-activity trends for the data set as a whole. The drawbacks related to the adoption of LLR are related to the efficiency of the determination of the local neighborhood, since all of the computations are done at query time. Moreover, uncorrelated features might result in errors in the identification of near neighbors. Finally, it is nontrivial to integrate feature selection in this framework. LLR is generally used to develop linear models for data sets in which the global structure-activity/property relationship is nonlinear in nature, as shown in the some recent works.[199–201]

III.2.4 CONTINUUM SOLVATION MODEL (CSM)

A continuum model in computational molecular sciences is defined as a model in which a number of the degrees of freedom of the constituent particles (a large number, indeed) are described continuously, usually by means of a distribution function. This principles has been adopted to develop solvation dynamics models which examines time-dependent, non-equilibrium, relaxation of the solvent (considered as a continuous dielectric medium), due to an instantaneous change in the interactions between the solvent molecules and the solute one, characterized by a brand new electronic state and embedded within a void cavity shaped by each continuum model. Each molecule involved in the system is typified by its calculated conductor screening charge density, called σ, which is describing the distribution of charges at a Klamt radii distance from atomic nuclei, i.e. an empirical value at which iso-electron-density surface of the molecule is reached, and represents the extension of the cavity containing the molecule.[202] The solvent molecules surrounding the solute are, initially, equilibrated with the ground electronic state of the solute molecule. As soon as the solute molecule are excited, the solvent ones move to a non-equilibrium state and need to adjust their positions and momenta in order to relax to equilibrium with the new interactions associated to the new electronic state. This relaxation is driven by a decrease in the solvent system interaction energy, which can be employed as an indicator for the underlying solvation dynamics of the system.[203]

Among the numerous models available, the COnductor like Screening MOdel (COSMO) proposed by Klamt and Schüürmann[204] attained a wide acceptance, since it can be applied to large and irregularly shaped molecular structures, requires a low numerical computational effort, produces accurate results for water and high permittivity solvents, and reduces the artifacts caused by small part of the electron density reaching outside of the cavity (outlying charge errors) in comparison to other dielectric continuum methods.[204] Klamt further developed this method in order to distinguish between two solvents with essentially identical dielectric constants, implementing statistical thermodynamics treatment of interacting surfaces in his late methodology, called conductor-like screening model for real solvents (COSMO-RS).[204]

The theory behind this extension of the COSMO model describes the interactions in a fluid as local contact interactions of molecular surfaces, whose interaction energies are quantified using the screening charge densities, called σ and σ' surfaces, that form a molecular contact between a couple of adjacent molecules. This assumption leads to a different representation of the solute/solvent system, since the ensemble of interacting molecules can be treated as an ensemble of independently interacting surface segments. The computation of the interaction energies between pairwise molecules embeds terms related to various contributions:

- *Electrostatics*, $e_{misfit}(\sigma, \sigma')$: the quantification of this interaction contemplates two scenarios. The former emerges when σ =−σ', as, due to a perfect fit of the two polarities, the solvent and the solute are still characterized by the same screening charge density as it was before the contact, hence, the electrostatic energy change would be ideally zero. The latter probes the presence of a local misfit interaction of the two polarization charge densities, which increases squared with the net charge.
- *Hydrogen Bonds*, $e_{HB}(\sigma, \sigma')$: this contribution is not null if one of the segments of the contact is an hydrogen bond donor and the other is an acceptor, providing a significant contribution to the real interaction of the two molecules.
- *Van der Waals:* the solvent, seen as a continuous conductor, also behaves like an average van der Waals partner, i.e. it makes dispersive interactions with the solute, which at a first approximation can be described as surface proportional contributions with element-specific coefficient.

For the calculation of each term, the reader can find more information on Klamt manuscript.[202]

The final expression describing the interaction energies of molecules, ΔE_{int}, treated as surface contact interactions and quantified by the local conductor polarization charge densities σ and σ' of the surface contacting segments, $a_{contact}$, is reported underneath (Eq.40):

$$\Delta E_{int} \cong a_{contact} e_{int}(\sigma, \sigma') = a_{contact}\left[e_{misfit}(\sigma, \sigma') + e_{HB}(\sigma, \sigma')\right] \qquad Eq.40$$

Which is then converted from surface contact interactions to fluid phase thermodynamics of a liquid ensemble S introducing the sum over all surface contacts k of the molecules belonging to the ensemble (Eq.41):

$$E_{tot}^{S} \cong \sum_{k} a_k e_{int}(\sigma_k, \sigma_k') \cong \sum_{k} a_{eff} e_{int}(\sigma_k, \sigma_k') \qquad Eq.41$$

where the effective contact area, a_{eff}, is the average contact area of the pair contacts k. Since the calculation of Eq.41 requires the computation of all contacts k and the relative polarization charge densities σ_k and σ_k', which are subject to permanent fluctuation in a liquid system, Klamt introduced statistical thermodynamics, ending up to reformulate the problem in favor of finding the σ-potential $\mu_s(\sigma)$, i.e. the chemical potential of an effective surface segment of area a_{eff} and polarity σ in the ensemble S. The chemical potential, which is depending on the interaction energies, the effective area and the temperature of the system, has a strong relationship with Kow following Eq.42, which is the equation adopted by Klamt for Kow estimation:

$$\log Kow = \log\left(exp\left[\frac{\left(\mu_j^w - \mu_j^o\right)}{RT} \left(\frac{V_o}{V_w}\right) \right] \right) \qquad Eq.42$$

in which μ_j^w is the chemical potential of solute j in water, μ_j^o represents the chemical potential of same solute in octanol and V_o and V_w are, respectively, the volumes of octanol and water in the system.

COSMO-RS allows for the prediction of all kinds of thermodynamic equilibrium properties of liquids, including solvation energies, activity coefficients and vapor pressure. Unlike group contribution methods, whose models rely on an extremely large number of experimental data, COSMO-RS calculates the thermodynamic data from molecular surface polarity distributions, which result from quantum chemical calculations of the individual compounds in the mixture using Density Functional Theory (DFT), extensively reviewed in Geerlins et al.[205] Using an efficient thermodynamic solution for such pairwise surface interactions, COSMO-RS converts the molecular polarity information into standard thermodynamic data of fluids.[203]

III.3 EVALUATION OF DIFFERENT KOW ESTIMATION METHODS APPLIED ON PHARMACEUTICAL COMPOUNDS

Numerous comparison of the performances of different Kow estimation methods have been published so far, commonly applying various QSAR methodologies to specific chemical groups.[206,207] Since the methodology presented in this thesis is focused on the sustainability assessment of fine chemical processes, a short survey on different available techniques for Kow estimation has been performed on common pharmaceutical compounds, aiming to identify the most suitable method in this field.

Thirty-two Active Pharmaceutical Ingredients (API), displayed in Table III-1 , have been chosen from different classes, comprehending antibiotics, a calcium channel blocking agent, a platelet aggregation inhibitor, a cardioselective beta blocker, antihistamines, nonsteroidal anti-inflammatory agents, a vitamin, a diuretic, a contrast media, an antineoplastics, antivirals, a thyroid drug, an antidepressant, a dermatological agent, proton pump inhibitors, analgesics, an anticonvulsant, and enzyme inhibitors. For these chemicals, experimental Kow values have been retrieved from webpage Drugbank[208] or, in case of missing data, from Chemspider[209] or Pubchem.[210] Experimental values have been compared to in-silico estimations using methods belonging to CSM group (COSMO-RS) and QSAR techniques.

Table III-1. Pharmaceutical compounds included in the analysis

	Chemical name	***CAS***	***FORMULA***	***Mol. Weight***	***Kow***
1	Amlodipine	88150-42-9	$C_{20}H_{25}ClN_2O_5$	408.879	3.00
2	Amoxicillin	26787-78-0	$C_{16}H_{19}N_3O_5S$	365.40	0.87
3	Aspirin	50-78-2	$C_9H_8O_4$	180.16	1.19
4	Atenolol	29122-68-7	$C_{14}H_{22}N_2O_3$	266.34	0.16
5	Atorvastatin	134523-00-5	$C_{33}H_{35}FN_2O_5$	558.64	5.70
6	Azelastine	58581-89-8	$C_{22}H_{24}ClN_3O$	381.90	4.90
7	Carbamazepine	298-46-4	$C_{15}H_{12}N_2O$	236.27	2.45
8	Cetirizine	83881-51-0	$C_{21}H_{25}ClN_2O_3$	388.89	2.80
9	Clavulanic Acid	58001-44-8	$C_8H_9NO_5$	199.16	-1.50
10	Diclofenac	15307-86-5	$C_{14}H_{11}Cl_2NO_2$	296.15	4.51
11	Ergocalciferol	50-14-6	$C_{28}H_{44}O$	396.65	7.30
12	Furosemide	54-31-9	$C_{12}H_{11}ClN_2O_5S$	330.75	2.03
13	Hydrochlorothiazide	58-93-5	$C_7H_8ClN_3O_4S_2$	297.74	-0.07
14	Ibuprofen	15687-27-1	$C_{13}H_{18}O_2$	206.29	3.97
15	Iopamidol	60166-93-0	$C_{17}H_{22}I_3N_3O_8$	777.08	-2.42
16	Irinotecan	100286-90-6	$C_{33}H_{38}N_4O_6$	586.68	3.20
17	Ketoprofen	22071-15-4	$C_{16}H_{14}O_3$	254.28	3.12
18	Ledipasvir	1256388-51-8	$C_{49}H_{54}F_2N_8O_6$	889.00	3.80
19	Lenalidomide	191732-72-6	$C_{13}H_{13}N_3O_3$	259.26	-0.40
20	Levothyroxine	51-48-9	$C_{15}H_{11}I_4NO_4$	776.87	4.00
21	Metoprolol	51384-51-1	$C_{15}H_{25}NO_3$	267.36	1.88
22	Naproxen	22204-53-1	$C_{14}H_{14}O_3$	230.26	3.18
23	Nortriptyline	72-69-5	$C_{19}H_{21}N$	263.38	4.51
24	Omeprazole	73590-58-6	$C_{17}H_{19}N_3O_3S$	345.42	2.20
25	Pantoprazole	102625-70-7	$C_{16}H_{15}F_2N_3O_4S$	383.37	0.50
26	Paracetamol	103-90-2	$C_8H_9NO_2$	151.16	0.46
27	Phenoxymethylpenicillin	87-08-1	$C_{16}H_{18}N_2O_5S$	350.39	2.09
28	Phenytoin	57-41-0	$C_{15}H_{12}N_2O_2$	252.27	2.47
29	Pregabalin	148553-50-8	$C_8H_{17}NO_2$	159.23	-1.35
30	Ramipril	87333-19-5	$C_{23}H_{32}N_2O_5$	416.51	2.90
31	Rosuvastatin	287714-41-4	$C_{22}H_{28}FN_3O_6S$	481.54	0.13
32	Sofosbuvir	1190307-88-0	$C_{22}H_{29}FN_3O_9P$	529.45	1.62

III.3.1 COSMO-RS METHODOLOGY PERFORMANCE

COSMO-RS methodology[202] has been used for Kow estimation using a quantum-mechanical CSM approach. This technique requires the adoption of three different software: the first provides a platform where is possible to build up the molecular structures and optimize their geometry finding the most stable ones, characterized by the minimum energy content. In this work Materials Studio by Accelrys has been used for this purpose, since it is able to predict the energy level of the molecules in the ideal gas phase using numerous force fields, i.e. a functional form and parameter sets employed to quantify the potential energy of a system of atoms. Hence, after the construction of the molecular structures, the geometry optimization steps have been performed using two distinct force fields, i.e. PCFF and COMPASS, resulting in different molecular shapes.

The next step contemplates the quantum-mechanical calculations on each optimized molecular structure, aiming to generate the σ-surfaces related to each chemical, which usually takes the major computational effort and the longest duration. After a conversion of the output file to a three dimensional coordinate file extension .sdf, the minimized molecular structures have been imported in TURBOMOLE in order to compute the quantum-mechanical calculations. During this step, two more configurations have been added to the evaluation, the former estimated using a TURBOMOLE original geometry optimization procedure and the latter retrieved from Pubchem database,[210] whose structures have already been minimized using MMFF94s force field.[211] The calculation steps require the selection of the methodology to pursue (in this work the BP-TZVP-COSMO has been chosen), the estimation of molecular orbital energies comprehending HOMO (highest occupied molecular orbital) and LUMO (lowest unoccupied molecular orbital), the level of theory to adopt (in this analysis, DFT has been employed) and definition of the accuracy, number of iterations and tolerance of the calculation. The details regarding the script elements available in TURBOMOLE are reported in Table III-2, which has been adapted from Toma.[23]

Table III-2. Details of TURBOMOLE Calculate script.

SYNTAX	DETAILS	
-l <list of molecules>	The list contains one or more molecules and its/their charge. The names of the molecule have to be equal to the input file names, without the extension (.<filetype>)	
-m <method>	Implemented methods	
	BP-TZVP-GAS	RI-DFT gas phase geometry optimization utilizing the b-p functional and def-TZVP basis set with standard settings
	BP-TZVP-COSMO	RI-DFT COSMO geometry optimization similar to BP-TZVP-GAS
	BP-SVP-GAS	RI-DFT gas phase geometry optimization utilizing the b-p functional and def-svp basis set with standard settings
	BP-SVP-COSMO	RI-DFT COSMO geometry optimization similar to BP-SVP-GAS
	BP-SVP-GAS-SP	Single point calculation with the same settings as BP-SVP-GAS
	BP-SVP-COSMO-SP	Single point calculation with the same settings as BP-SVP-COSMO
	AM1-GAS	MOPAC AM1 gas phase geometry optimization
	AM1-COSMO	MOPAC AM1 COSMO geometry optimization
-f <filetype>	Implemented input files types	
	car	Biosym car files
	ML2	TRIPOS mol2 files
	arc	MOPAC archive files
	cosmo	Cosmo files (for recalculation)
	xyz	XYZ format: 1) number of atoms 2) comment line 3) symbol XYZ(for each atom)
-din <dir>	Input file (3D coordinate files) directory	
-dcos <dir>	cosmo/energy/arc file directory (to collect cosmo, energy or MOPAC archive files)	
-dcomp <dir>	Parent directory for TURBOMOLE calculation	

The output files produced with extension .cosmo have been exported to COSMOtherm software in order to gain the estimation of Kow. Each configuration has been evaluated individually, comparing the in-silico estimated Kow value with the experimental one. Moreover, the σ-profiles of the four configurations have been merged into a single entity, taking into consideration the contributions expressed by all four simultaneously, producing a fifth input to assess, which has been called “Merged” further on. The results obtained from the Kow estimation procedures for the four different configurations (TURBOMOLE, MMFF94s, COMPASS and PCFF) and the “Merged” one have been reported in Table III-3. The comparison between each set of data with the experimental values has been performed through the generation of linear regression models using the R function <lm> within the Rstudio platform. The <lm> function couples the experimental values with the predicted ones, using the following Eq.43:

$$y_{exp} = a + b \cdot x_{pred} \qquad Eq.43$$

in which y_{exp} and x_{pred} represent the experimental and the predicted values, respectively, while a and b are the parameters (intercept and slope) of the specific model regressed.

Table III-3. Result of different Kow estimation methodologies using various minimization techniques

		Experimental	***TURBOMOLE***	***MMFF94s***	***COMPASS***	***PCFF***	***Merged***
1	Amlodipine	3	3.53	3.64	2.76	2.05	3.59
2	Amoxicillin	0.87	0.45	0.60	0.58	1.21	0.45
3	Aspirin	1.19	1.75	1.75	1.22	1.14	1.75
4	Atenolol	0.16	1.70	0.23	2.14	2.50	1.62
5	Atorvastatin	5.7	3.45	4.74	4.78	5.78	4.74
6	Azelastine	4.9	5.21	4.92	5.74	5.90	5.00
7	Carbamazepine	2.45	1.43	1.43	1.57	2.25	1.43
8	Cetirizine	2.8	4.68	4.80	5.28	5.78	4.80
9	Clavulanic_Acid	-1.5	-0.64	-0.12	0.05	0.32	-0.13
10	Diclofenac	4.51	2.96	4.47	4.52	4.20	4.47
11	Ergocalciferol	7.3	8.48	8.86	8.33	8.68	8.86
12	Furosemide	2.03	2.41	2.41	1.44	2.20	2.41
13	Hydrochlorothiazide	-0.07	-0.84	-0.70	0.11	-0.37	-0.83
14	Ibuprofen	3.97	3.97	3.88	2.92	3.36	3.94
15	Iopamidol	-2.42	-2.67	-2.67	-1.16	-0.57	-2.67
16	Irinotecan	3.2	5.61	5.51	5.74	6.20	5.51
17	Ketoprofen	3.12	2.97	2.84	2.57	3.21	2.89
18	Ledipasvir	3.8	6.29	4.78	7.44	8.39	6.29
19	Lenalidomide	-0.4	-0.71	-1.23	-1.32	-0.89	-1.23
20	Levothyroxine	4	3.93	4.83	4.77	5.16	4.81
21	Metoprolol	1.88	4.13	3.13	4.11	4.14	3.72
22	Naproxen	3.18	3.15	3.13	2.21	2.63	3.14
23	Nortriptyline	4.51	4.98	4.95	4.70	4.74	4.95
24	Omeprazole	2.2	2.21	2.07	2.96	1.99	2.07
25	Pantoprazole	0.5	2.34	2.22	2.89	2.22	2.24
26	Paracetamol	0.46	0.29	0.27	0.49	0.56	0.27
27	Phenoxymethylpenicillin	2.09	-1.17	1.81	1.80	2.77	1.81
28	Phenytoin	2.47	1.29	1.20	1.14	1.51	1.24
29	Pregabalin	-1.35	1.04	1.41	1.47	1.47	1.41
30	Ramipril	2.9	2.84	4.16	4.50	6.15	4.16
31	Rosuvastatin	0.13	1.74	2.45	4.24	3.66	2.44
32	Sofosbuvir	1.62	1.62	1.39	2.34	3.69	1.39

Beside the five prediction model, a backward elimination procedure has been performed, aiming to identify the best possible combination of each model contribution simultaneously in order to enhance the overall regression performance. The procedure is well explained in a Crawley's book[212] and consists

in the generation of a model that embeds all the possible combinations of every model contribution, comprehending single linear, individual squared and multiple-way interactions. Then simplification model process begins, eliminating stepwise the contributions that are clearly not significant (considering the p-value). This refining process has been carried on until each contribution was deemed significant and a further simplification would have worsen the overall model performance. The performances of the models are reported in Table III-4, in which the best model, i.e. the backward elimination, is highlight by the highest value of r^2=0.8172. However, it is worth underlying how the performance of MMFF94s is comparable to the best one, since its r^2 of 0.8030 is rather close to the optimal value. Furthermore, the prediction of Kow using a single optimized configuration, as in MMFF94s, is preferable, as it enables to save a considerable amount of time and computational effort.

Table III-4. Results of COSMO evaluation models

	COSMO MODELS					
	TURBOMOLE	*MMFF94s*	*COMPASS*	*PCFF*	*Merged*	*Backward Elimination*
Residual Standard Error	1.253	0.9853	1.372	1.346	1.049	0.949
Adjusted R-squared, R^2	0.6816	0.8030	0.6182	0.6322	0.7768	0.8172
F statistics	67.36	127.4	51.2	54.28	108.9	20.8
p-value	$3.67 \cdot 10^{-09}$	$2.55 \cdot 10^{-12}$	$5.83 \cdot 10^{-08}$	$3.31 \cdot 10^{-08}$	$1.68 \cdot 10^{-11}$	$9.55 \cdot 10^{-09}$

The coefficients related to the backward elimination model are shown in Table III-5, including the standard error and the p-value (Pr) related to each coefficient that multiplies the contribution of each significant term. The relevancy of each term is indeed guaranteed by the low values of the p-values of the last column.

Table III-5. Coefficients of backward elimination model terms

	Coefficient	**Std. Error**	**t value**	**Pr(>|t|)**
(Intercept)	-0.17879	0.29032	-0.616	0.5438
TURBOMOLE	0.45649	0.15172	3.009	0.006077
MMFF94s:Merged	0.2803	0.09456	2.964	0.006753
Merged:COMPASS	-1.03159	0.25011	-4.124	0.000385
Merged:PCFF	0.90413	0.2234	4.047	0.000468
TURBOMOLE:MMFF94s:COMPASS	-0.07807	0.03373	-2.315	0.029509
TURBOMOLE:Merged:COMPASS	0.34309	0.10113	3.392	0.002402
TURBOMOLE:Merged:PCFF	-0.27261	0.07724	-3.53	0.001712

The plots shown in Figure III-6 have been produced using RStudio in order to ensure the good quality of the backward elimination model, as explained in Crawley.[212] The left plot of the figure shows the residuals on the *y* axis against fitted values on the *x* axis, thus it is desirable not to recognize any structure or specific pattern in the plot to avoid the inconstancy of variance and this is the case of the chart obtained. Meanwhile, Figure III-6 right plot represents the normal quantile–quantile plot, which should be a straight line if the errors are normally distributed, as S-shaped or banana-shaped pattern would need a different model to fit the data. Again the chart obtained is quite satisfactory in term of distribution of errors.

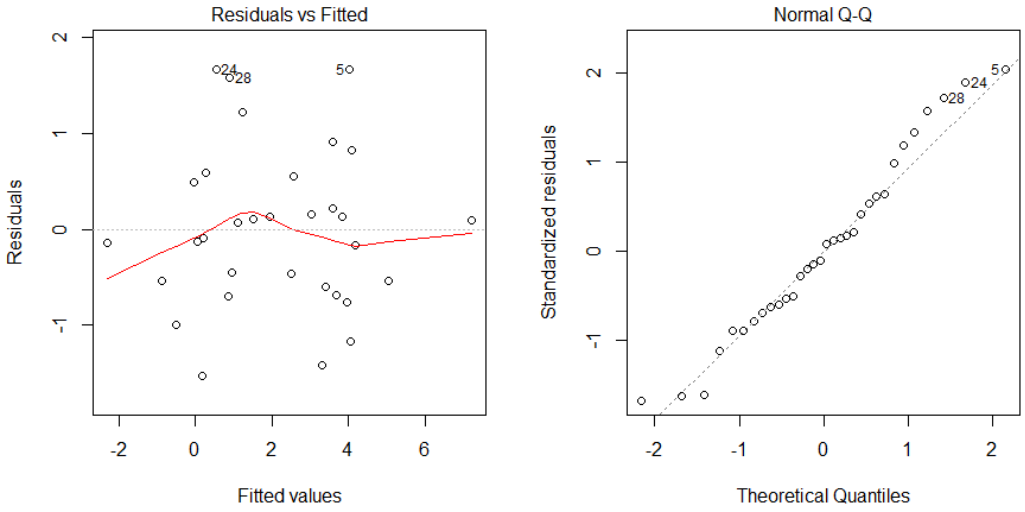

Figure III-6. Plots for evaluation of backward elimination model

An analogous procedure has been performed to ensure the good quality of MMFF94s-based prediction. The left plot of Figure III-7 shows how the residuals on the *y* axis against the fitted values on the *x* axis, are randomly distributed, ensuring the constancy of variance, as demanded. Nevertheless, Figure III-7 right plot, which represents the normal quantile–quantile plot, displays a non-ideal residuals trend for several values, which stand rather far from the linear trendline, due to some outliers whose values is poorly predicted by this estimation technique. Since in this work this set of Kow estimated values have been retrieved within Pubchem database,[210] the original minimization software has not been used, thus it has not been possible to detect the issues related to the poor estimation results. However, further studies employing the original force field have been already planned in the near future.

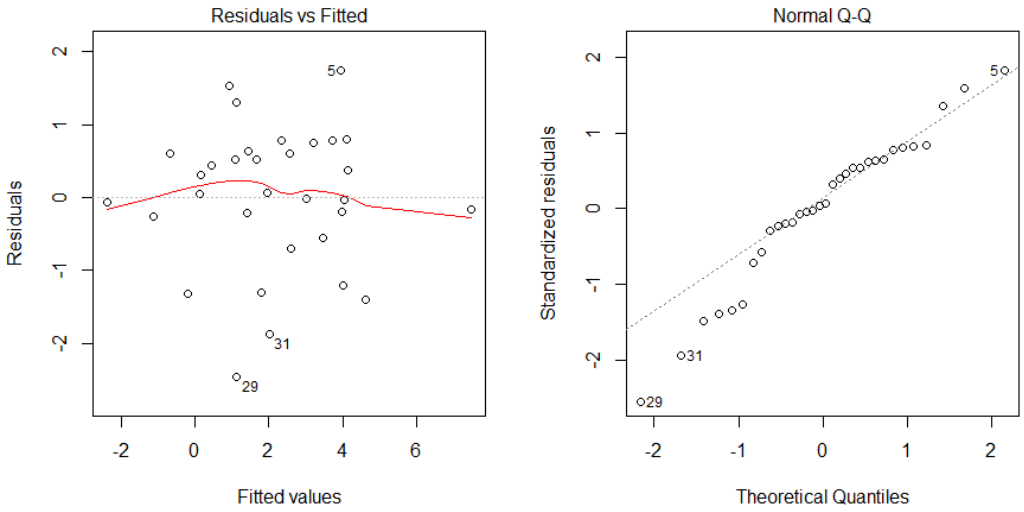

Figure III-7.Plots for evaluation of MMFF94s-based model

III.3.2 QSAR METHODOLOGIES PERFORMANCE

The performances of well-established QSAR methodologies for Kow estimation have been investigated, aiming to identify the most promising one in the field of pharmaceutical compounds, evaluating the same thirty-two chemicals from different drug classes.

Since QSAR techniques usually don't require high computational effort, some methodologies have already been embedded within some webpages, where predicted values can be easily retrieved. Indeed, Kow values estimated using Percepta ACD/Lab have been retrieved on Chemspider,[209] ALOGPS and ChemAxon results are available on Drugbank,[208] Pubchem[210] provides estimations from XlogP3, and Toxcast[213] evaluates Kow values using QikProp. Furthermore, KOWWIN platform, part of EPI Suite software downloaded from EPA,[120] has been adopted for genuine Kow prediction, using SMILES as input. As a result, six new Kow evaluations have been produced and a comparison among them has been performed.

The results of the Kow evaluations are reported in Table III-6, which displays the results obtained using the different QSAR techniques adopted.

Table III-6. Result of different Kow estimation methodologies comprehending various QSAR approaches

			QSAR Methodologies					
	Chemical name	*Experimental*	*Kowwin*	*ACD/Lab*	*ALOGPS*	*ChemAxon*	*XlogP3*	*QikProp*
Source	*Scifinder*	*Drugbank*	*EPISuite*	*Chemspider*	*Drugbank*	*Drugbank*	*Pubchem*	*EPA iCSS Toxcast*
1	Amlodipine	3	2.07	4.16	2.22	1.64	3.00	-
2	Amoxicillin	0.87	0.97	0.92	0.75	-2.30	-2.00	-2.25
3	Aspirin	1.19	1.13	1.19	1.43	1.24	1.20	1.19
4	Atenolol	0.16	-0.03	0.10	0.57	0.43	0.20	0.16
5	Atorvastatin	5.7	6.36	4.93	4.41	5.39	5.00	6.57
6	Azelastine	4.9	5.72	3.71	3.81	4.04	4.40	-
7	Carbamazepine	2.45	2.25	2.67	2.10	2.77	2.50	2.39
8	Cetirizine	2.8	-0.61	2.16	2.98	0.86	1.70	-
9	Clavulanic Acid	-1.5	-2.04	-1.98	-1.20	-1.50	-1.20	-
10	Diclofenac	4.51	3.92	4.06	4.98	4.26	4.40	4.50
11	Ergocalciferol	7.3	10.44	9.56	7.59	7.05	7.40	7.32
12	Furosemide	2.03	2.32	3.10	2.71	1.75	2.00	1.83
13	Hydrochlorothiazide	-0.07	-0.10	-0.07	-0.16	-0.58	-0.10	-0.07
14	Ibuprofen	3.97	3.79	3.72	3.50	3.84	3.50	3.50
15	Iopamidol	-2.42	0.23	-2.09	-0.97	1.62	-2.40	-0.85
16	Irinotecan	3.2	2.33	4.35	3.94	2.78	3.00	3.35
17	Ketoprofen	3.12	3.00	2.81	3.29	3.61	3.10	3.12
18	Ledipasvir	3.8	6.89	6.77	5.92	6.11	7.40	-
19	Lenalidomide	-0.4	-1.99	-1.39	-0.43	-0.71	-0.50	0.06
20	Levothyroxine	4	4.12	5.93	1.15	3.73	2.40	2.07
21	Metoprolol	1.88	1.69	1.79	1.80	1.76	1.90	1.89
22	Naproxen	3.18	3.10	3.00	3.29	2.99	3.30	3.18
23	Nortriptyline	4.51	4.74	5.65	4.65	4.43	4.50	-
24	Omeprazole	2.2	3.40	2.17	1.66	2.43	2.20	2.23
25	Pantoprazole	0.5	2.22	1.69	2.11	2.18	2.40	2.65
26	Paracetamol	0.46	0.27	0.34	0.51	0.91	0.50	0.51
27	Phenoxymethylpenicillin	2.09	1.87	1.88	1.78	0.76	2.10	-
28	Phenytoin	2.47	2.16	2.29	2.26	2.15	2.50	2.47
29	Pregabalin	-1.35	-1.78	1.12	-1.40	-1.30	-1.60	-1.30
30	Ramipril	2.9	3.32	3.41	0.92	1.47	1.40	1.02
31	Rosuvastatin	0.13	2.48	0.42	1.47	1.92	1.60	2.91
32	Sofosbuvir	1.62	0.75	1.62	1.63	1.28	1.00	0.92

The comparison between each set of estimated data with the experimental values has been performed as for COSMO models, i.e. through the generation of regression models using the R function <lm> within the Rstudio platform. The performances of the models are reported in Table III-7, from which the best model arise as the ACD/Labs used by Chemspider, since is characterized by the highest value of r^2=0.8497 and the lowest p-value.

Table III-7. Performances of QSAR models involved in the Kow estimation analysis

	QSAR Models					
	Kowwin	*ACD/Lab*	*ALOGPS*	*ChemAxon*	*XlogP3*	*QikProp*
Residual standard error	1.089	0.8749	0.9962	1.253	1.021	1.204
Adjusted R-squared, R^2	0.7595	0.8447	0.7986	0.6813	0.7885	0.7135
F statistic	98.91	169.6	123.9	67.26	116.6	60.76
p-value	$5.20 \cdot 10^{-11}$	$7.03 \cdot 10^{-14}$	$3.55 \cdot 10^{-12}$	$3.73 \cdot 10^{-09}$	$7.45 \cdot 10^{-12}$	$6.68 \cdot 10^{-08}$

Following the prescriptions of Crawley,[212] the plots of Figure III-8 have been generated using RStudio in order to ensure the good quality of the model.

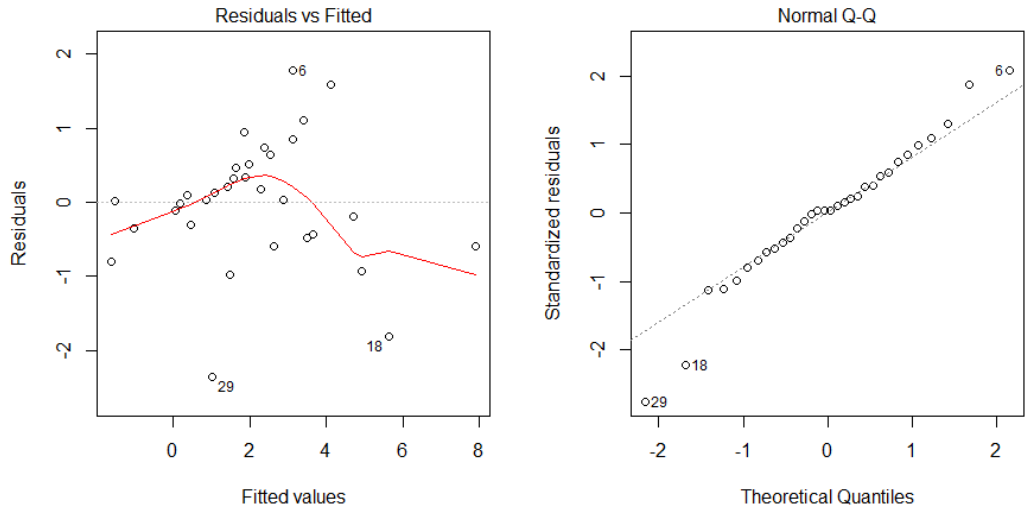

Figure III-8.Plots for evaluation of ACD Percepta model

The left plot of Figure III-8 shows the residuals on the *y* axis against fitted values on the *x* axis, thus it is desirable not to recognize any structure or specific pattern in the plot to avoid the inconstancy of variance. Unluckily, the residuals are not equally distributed, supporting the hypothesis of the adoption of a response variable different than linear. On the contrary, Figure III-8 right plot, which represents the normal quantile–quantile plot, shows that ACD model is quite reliable, as it efficiently follows the ideal distribution of errors. However, some estimated values are still far from a good estimation: Kow of Ledipasvir (ID 18) is probably overestimated since it is composed by a lot of groups that contribute to the final Kow value, Pregabalin (ID 29) is quite controversy, since the other QSAR models estimate its Kow more closely than ACD. Since in this thesis the original software has not been adopted, it has not been possible to detect the issues related to these estimated value. In this perspective, an increased number of pharmaceutical compounds will be assessed in further studies, aiming to identify the best Kow estimation model, using the original software.

III.3.3 IDENTIFICATION OF THE MOST ACCURATE METHODOLOGY

The adequate conclusion of this short analysis resides on a comparison between the best models highlighted by quantum mechanical and QSAR approaches with the purpose of identifying the most reliable estimation technique for the compounds under investigation.

Table III-8 shows the experimental and the predicted values for the most accurate methodology from both approaches.

Table III-8. Experimental vs predicted values for the most accurate QSAR methodology and COSMO geometries

			QSAR	***COSMO***	
		Exp	*ACD/Lab*	*MMFF94s*	*Backward Elimination*
1	Amlodipine	3.00	4.16	3.64	3.68
2	Amoxicillin	0.87	0.92	0.60	0.29
3	Aspirin	1.19	1.19	1.75	1.12
4	Atenolol	0.16	0.10	0.23	0.86
5	Atorvastatin	5.70	4.93	4.74	4.04
6	Azelastine	4.90	3.71	4.92	4.07
7	Carbamazepine	2.45	2.67	1.43	1.24
8	Cetirizine	2.80	2.16	4.80	3.40
9	Clavulanic_Acid	-1.50	-1.98	-0.12	-0.50
10	Diclofenac	4.51	4.06	4.47	3.60
11	Ergocalciferol	7.30	9.56	8.86	7.21
12	Furosemide	2.03	3.10	2.41	2.49
13	Hydrochlorothiazide	-0.07	-0.07	-0.70	0.06
14	Ibuprofen	3.97	3.72	3.88	3.85
15	Iopamidol	-2.42	-2.09	-2.67	-2.28
16	Irinotecan	3.20	4.35	5.51	3.95
17	Ketoprofen	3.12	2.81	2.84	2.57
18	Ledipasvir	3.80	6.77	4.78	3.59
19	Lenalidomide	-0.40	-1.39	-1.23	-0.86
20	Levothyroxine	4.00	5.93	4.83	4.17
21	Metoprolol	1.88	1.79	3.13	3.30
22	Naproxen	3.18	3.00	3.13	3.03
23	Nortriptyline	4.51	5.65	4.95	5.05
24	Omeprazole	2.20	2.17	2.07	0.54
25	Pantoprazole	0.50	1.69	2.22	0.95
26	Paracetamol	0.46	0.34	0.27	-0.02
27	Phenoxymethylpenicillin	2.09	1.88	1.81	1.96
28	Phenytoin	2.47	2.29	1.20	0.90
29	Pregabalin	-1.35	1.12	1.41	0.58
30	Ramipril	2.90	3.41	4.16	4.06
31	Rosuvastatin	0.13	0.42	2.45	0.22
32	Sofosbuvir	1.62	1.62	1.39	1.51
	Adjusted R-squared		0.843	0.799	0.817
	AIC		86.19	93.80	96.26

The criteria employed for the evaluation of models performances have been the Adjusted R-squared and the Akaike's Information Criterion (AIC). The Adjusted R-squared, R^2_{adj}, which is based on Eq.44, is a modification of the original R-squared, which explains the degree to which the input variables (i.e. the predicted values) explain the variation of the experimental ones. However, in a Multivariate Linear Regression, the inclusion of new variables induces an increase in the R-squared value, disregarding whether the addition concerns a significant variable. The Adjusted R-squared, instead, provide information on the percentage of variation explained by the fraction of input variables that actually describe the trend of the output ones, penalizing for adding meaningless independent variables (k in Eq.44) that do not fit the experimental values trend. Indeed, a different number of inputs has been employed for each model (one for ACD and MMFF94s, seven for Backward Elimination), therefore the adoption of Adjusted R-squared has been essential in order to avoid misleading conclusions.

$$R^2_{adj} = 1 - \left[\frac{(1 - R^2)(n - 1)}{n - k - 1}\right] \qquad Eq.44$$

where R^2 is the original R-squared value, n is the number of samples and k is the number of independent variables model.

The values of R^2_{adj} reported in Table III-8 do not show a relevant variation, thus another criterion has been introduced in order to provide a supplementary support to the model sorting, i.e. Akaike's Information Criterion (AIC).

This indicator has been calculated using an Rstudio implemented function called <AIC> which is based on Eq.45, in which n is the number of variables of the model and loglikelihood is the logarithmic function of the parameters of a statistical model given data.

$$AIC = -2loglikelihood + 2n \qquad Eq.45$$

The lower the value of AIC, the better the fit, therefore the QSAR ACD/Lab methodology arise as the best technique for the Kow prediction of the thirtytwo API under study, as shown in Table III-8.

In Chapter IV several case studies will be discusses with the purpose of showing the various field of application of the methodology. The first study will focus on a pharmaceutical application in which the production process of an API, *i.e.* pioglitazone hydrochloride, will be assessed for the sake of identifying the most sustainable routes among different process designs found in literature. The second case study will treat the production of biodiesel from palm oil, optimizing the reactor operating conditions from a sustainability viewpoint. The last application belongs to the nanotechnology field, i.e. the evaluation of the production of CdSe quantum dots using social, environmental and economic principles

CHAPTER IV

CASE STUDIES

IV.1 PIOGLITAZONE HYDROCHLORIDE PRODUCTION PROCESS

Hydrochloride salt of pioglitazone is a thiazolidinedione (TZD) class drug employed for type II diabetes treatment. It is produced by Takeda Pharmaceuticals under the brand name Actos and it is prescribed for its hypoglycemic action as a binder to the Peroxisome Proliferator-Activated Receptor Gamma (PPARγ).[214] Lipid and glucose homeostasis, as well as other various metabolic processes, are affected by PPARs which are ligand-activated transcription factors involved in the expression of a wide variety of genes.[215] Therefore, through its mechanism of action, pioglitazone hydrochloride improves utilization of glucose by increasing insulin sensitivity in adipose and muscle tissue, while reducing glucose production via the liver.

The epidemic diabetes data highlight an increasing trend on the spreading of this disease with an estimated rise of the global prevalence for all age groups from 2.8% (171 million people) in 2000 to 4.4% (366 million people) by 2030.[216] However, some studies revealed an association between the prolonged use of pioglitazone and the risk of developing bladder cancer.[217–219] Although the debate in the scientific community is still wide open due to other contradictory studies,[220–222] the increasing risk related to the administration of pioglitazone led to a massive decreasing in sales, from $2.8 billion in 2006[223] to $182 million in 2016,[224] and to a $2.37 billion compensation settlement for around ten thousands lawsuits in USA against Takeda Pharmaceuticals. Notwithstanding, this API is still produced and numerous synthesis routes can be retrieved in literature.

The sustainability evaluation methodology described in this thesis has been applied to several production processes, aiming to identify the most sustainable process alternative, following the scheme shown in Figure IV-1.

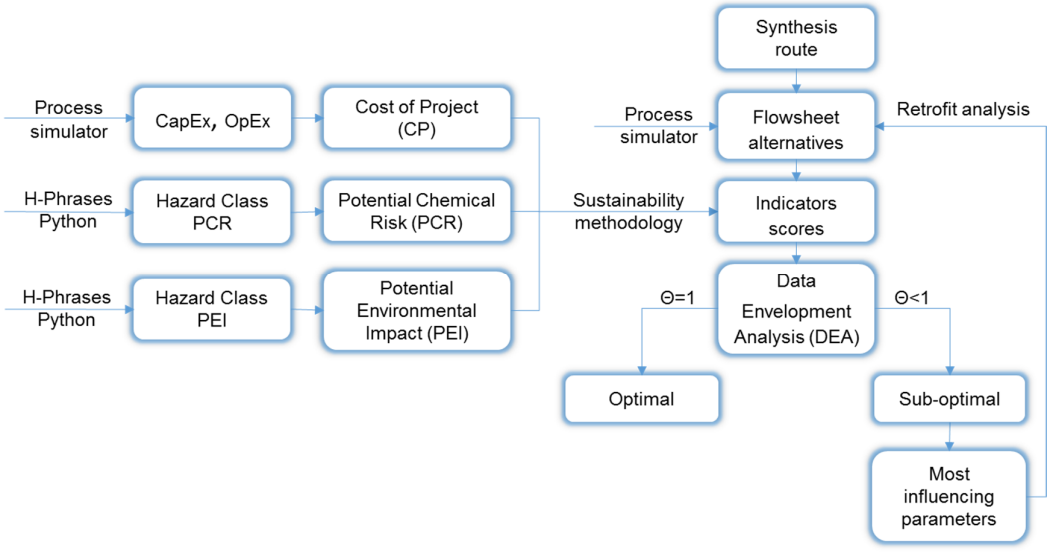

Figure IV-1. Decision path for sustainability assessment of pioglitazone hydrochloride production process

IV.1.1 RETRIEVING DIFFERENT ROUTES

Despite the issues related to Actos, pioglitazone is still under production and numerous patents with different routes and chemicals have been deposited.[223] In this dissertation, a selection of several available routes published in patents has been performed in order to highlight the different approaches among them, e.g. adoption of different solvents, reactants or catalysts. Some of the chemicals involved are referred using sequential numbers that have been reported in the scheme of the synthesis routes of Figure IV-2, Figure IV-3. The first published route selected is based on the work of Meguro et al.[225] and follows a five step procedure comprehending five reaction in series. At first, **1** reacts with **2** in an aqueous solution of benzyltributylammonium chloride, 1,2-dichloroethane and sodium hydroxide. The solution is then added with **8** and sodium hydroxide to obtain compound **13** which is extracted using 1,2-dichloroethane and water. Next step consists of the reaction of **13** with **15** in piperidine and ethanol, followed by extraction and recrystallization of product **17** using ethanol and 1,2-dichloroethane. Last step involves a hydrogenation of **17** catalyzed by palladium on carbon in N,N-dimethylformamide, followed by work-up using water and 1,4-dioxane to gain pioglitazone (**27**) as final product. In order to obtain the chlorinated salt of pioglitazone (**28**), a procedure from the subsequent paper of Sohda et al.[226] has been chosen. In Sohda's best proposed route, **1** reacts with **4** in a solution of N,N-dimethylformamide and NaH in oil, giving **14** as product. After an extraction with ethyl acetate, compound **14** is hydrogenated in methanol, adopting palladium on carbon as catalyst in order to obtain **18**. This reaction is followed by the addition to **18** of hydrogen bromide in an aqueous solution of $NaNO_2$ and then **22** is added coupled with Cu_2O as catalyst. In the next step, **25** reacts with **24** in ethanol and sodium acetate, followed by extraction of **26** with ethyl acetate. Last step involves the introduction of hydrochloric acid and ethanol in the reaction system in order to obtain pioglitazone hydrochloride (**28**) as final product. After ten years, Saito et al. registered a new patent[227] in which numerous routes were proposed, giving the opportunity to evaluate different designs. Among the alternatives, four individual routes have been included in this work, each one characterized by the utilization of different reactants or solvents. First route adopts toluene, triethylamine and methanesulfonyl chloride to perform the first reaction starting with **1** to obtain **6**, followed by addiction of **8**, ethanol, potassium carbonate, sodium hydroxide and activated carbon to give rise to **13**. Afterwards, compound **13**, **15**, ethanol and piperidine are mixed in order to gain product **17**, which is then charged to a high pressure and temperature hydrogenation reactor in presence of palladium on carbon and 1,4-dioxane. The product of this step is **27**, which is washed with 1,4-dioxane and ethanol and added with an aqueous solution of hydrochloric acid to obtain its chlorinated salt (**28**). The second route selected employs tetrahydrofuran, sodium hydroxide, and **2** for the first reaction with **1**, while the second step occurs in ethanol after an extraction of **13** using ethyl acetate. A third route follows the description of the first one, even though the reaction of **6** with **8** takes place in isopropanol, the subsequent one with **15** in methanol, pyrrolidine, triethylamine and hydrochloric acid

and the hydrogenation in tetrahydrofuran. The fourth route reflects the steps of the first one, although it involves both the utilization of dichloromethane as a solvent in the first reaction and the adoption of methanol, hydrochloric acid, tetrahydrofuran and sodium hydroxide in the hydrogenation step. Moving to a different source, Rajendra's proposed route[228] has been considered, as he adopts a novel procedure related to Momose's work[229] using p-nitrophenol (**9**), hydrobromic acid, methyl acrylate (**22**), a nickel-based and a copper-based catalysts (avoiding the purchasing of expensive palladium-based one), while it employs novel solvents, e.g. acetonitrile, acetone/water and diisopropylether. Suri's work[230] shares some steps with the former one, however it has been investigated due to the employment of water as a solvent in the reaction of **1** with **4** and diisopropyl ether and hexane as extraction agents. A substantial difference in the hydrogenation step can be retrieved in Madivada's work,[231] in which the utilization of cobalt chloride hexahydrate, sodium borohydride and dimethyl glyoxime as catalytic system[232] avoids the high pressure and temperature reaction previously adopted. The last three routes come from a combination of two patents,[233,234] in which the former describes the procedure to give rise to the starting material of the latter, i.e. 5-ethyl-2-vinyl-pyridine (**7**). In the first route a dissolution of **7** and **10** in tert-butanol occurs, followed by addition of aqueous sodium hydroxide and extraction by means of dichloromethane. Afterwards, **8** in toluene and aqueous sodium hydroxide are supplied to the reaction system, followed by work-up using PEG and diethyl ether. Subsequent steps involve **15** and pyrrolidine in methanol, hydrogenation catalyzed by cobalt chloride hexaydrate, zinc-catalyzed reaction with thionyl chloride in chloroform and final workup with hydrochloric acid in ethanol. The second route takes into account the replacement of tert-butanol with dimethylsulfoxide, while the third one introduces bromine and potassium bromide instead of **10** in the first step. The sources of the synthesis routes just mentioned have been summarized in Table IV-2, while Table IV-2 displays the substances involved in the study.

Table IV-1. Summary of synthesis routes selected

#	Route	Inventors	Steps
1	Flowsheet 1	Saito et al.[227]	Reference Example 3, Working Example 4, Reference Example 4, Reference Example 7, Reference Example 11
2	Flowsheet 1.1	Saito et al.[227]	Reference Example 2, Working Example 3, Reference Example 4, Reference Example 7, Reference Example 11
3	Flowsheet 1.2	Saito et al.[227]	Reference Example 5, Working Example 5, Reference Example 6, Reference Example 9, Reference Example 10
4	Flowsheet 1.3	Saito et al.[227]	Reference Example 1, Working Example 1, Reference Example 4, Reference Example 13, Reference Example 12
5	Flowsheet 2	Madivada et al.[231]	Experimental Section
6	Flowsheet 3	Meguro et al.[225] Saito et al.[227]	[Example 1-b), Example 2-c), Example 3-d)],[225] Reference Example 11[227]
7	Flowsheet 4	Mohanty et al.[233] Pandey et al.[234]	Experimental Section (3),[233] [Example 26, Example 36, Example 41, Example 51, Example 54, Example 68][234]
8	Flowsheet 4.1	Mohanty et al.[233] Pandey et al.[234]	Experimental Section (3),[233] [Example 2, Example 11, Example 26, Example 36, Example 41, Example 51, Example 54, Example 68][234]
9	Flowsheet 4.2	Mohanty et al.[233] Pandey et al.[234]	Experimental Section (3),[233] [Example 3, Example, 14, Example 26, Example 36, Example 41, Example 51, Example 54, Example 68][234]
10	Flowsheet 5	Saito et al.[227] Rajendra et al.[228]	Reference Example 3,[227] Experimental section[228]
11	Flowsheet 6	Sohda et al.[226] Oba et al.[235]	Experimental section,[226] Example 1[235]
12	Flowsheet 7	Suri et al.[230]	Example 1-a), Example 7-a), Example 9-b), Example 10, Example 11, Example 12

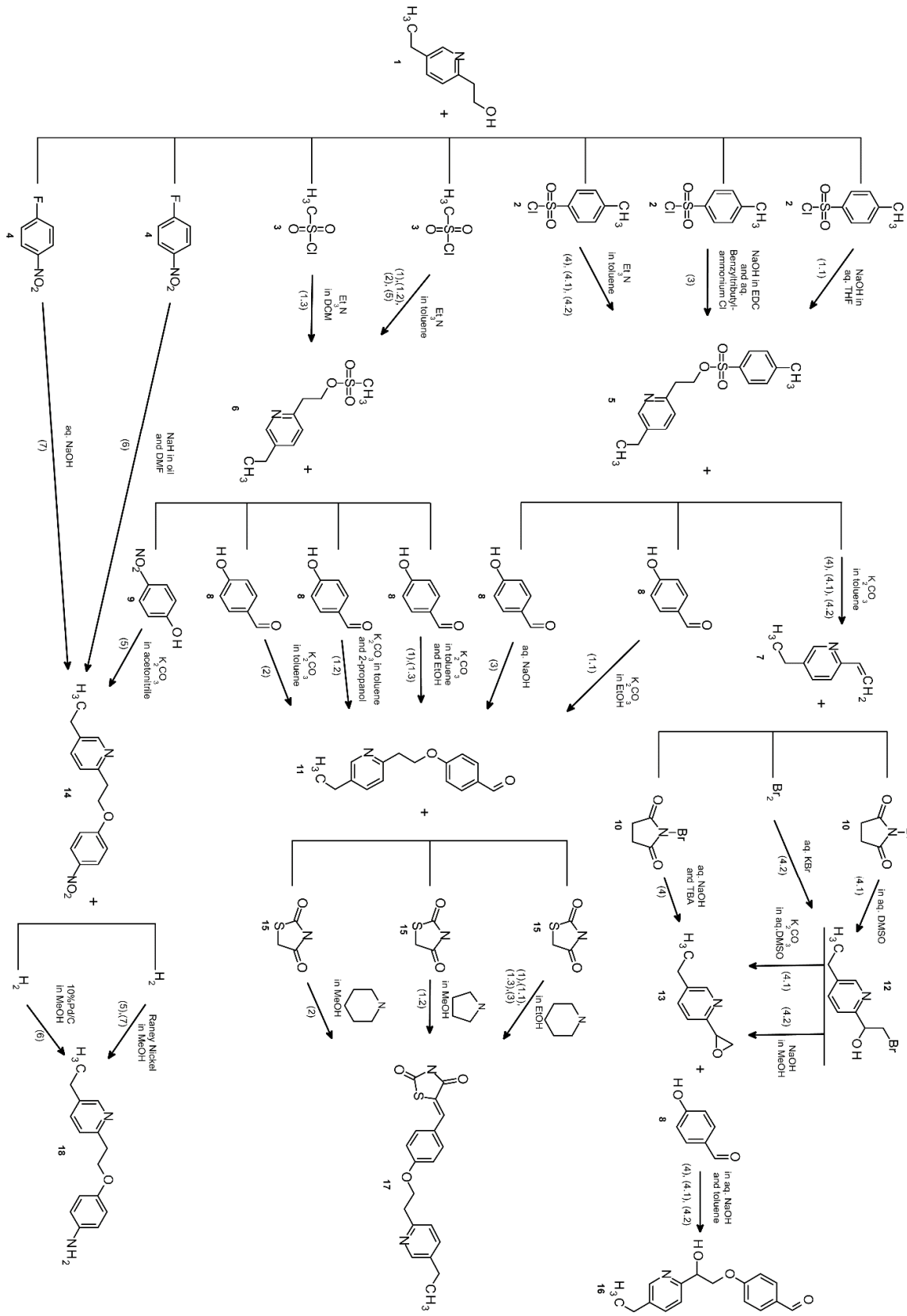

Figure IV-2. Scheme of the synthesis routes under study, which are specified under the reaction arrows in brackets – Part 1

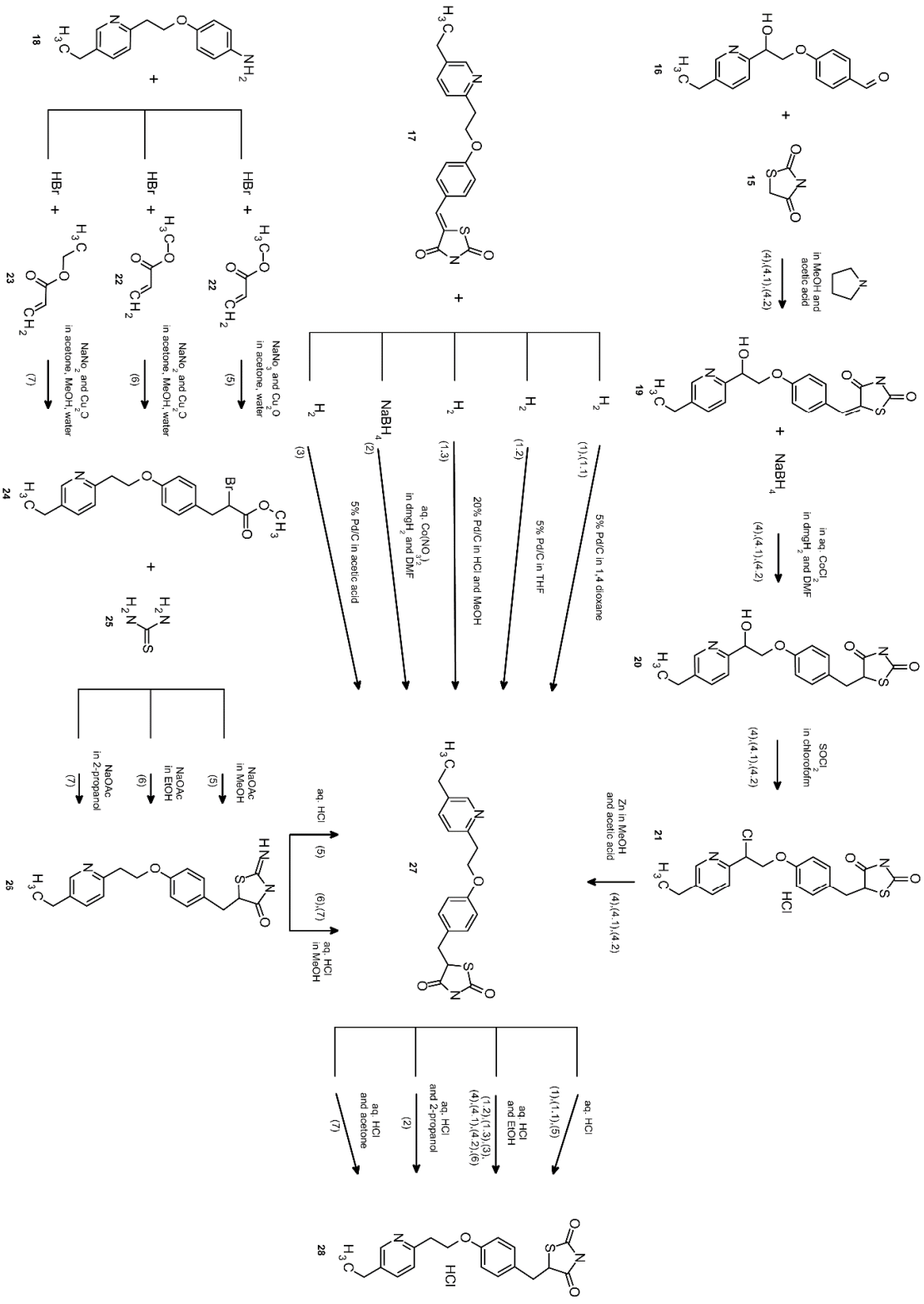

Figure IV-3. Scheme of the synthesis routes under study, which are specified under the reaction arrows in brackets – Part 2

Table IV-2. List of the substances involved in the synthesis routes under study

#	CAS	Name
1	5223-06-3	5-Ethyl-2-pyridineethanol
2	98-59-9	p-Toluenesulfonyl chloride
3	124-63-0	Methanesulfonyl chloride
4	350-46-9	p-Fluoronitrobenzene
5	144809-27-8	2-Pyridineethanol, 5-ethyl-, 4-methylbenzenesulfonate
6	144809-26-7	2-(5-Ethyl-2-pyridyl)ethyl methanesulfonate
7	5408-74-2	5-Ethyl-2-vinylpyridine
8	123-08-0	p-Hydroxybenzaldehyde
9	100-02-7	p-Nitrophenol
10	128-08-5	N-Bromosuccinimide
11	646519-81-5	α-(Bromomethyl)-5-ethyl-2-pyridinemethanol
12	471295-97-3	5-Ethyl-2-(2-oxiranyl)pyridine
13	114393-97-4	4-[2-(5-Ethyl-2-pyridinyl)ethoxy]benzaldehyde
14	85583-54-6	4-[2-(5-Ethyl-2-pyridinyl)ethoxy]nitrobenzene
15	2295-31-0	2,4-Thiazolidinedione
16	471295-98-4	4-[2-(5-Ethyl-2-pyridinyl)-2-hydroxyethoxy]benzaldehyde
17	144809-28-9	5-[4-[2-(5-Ethyl-2-pyridinyl)ethoxy]benzylidene]-1,3-thiazolidine-2,4-dione
18	85583-40-0	4-[2-(5-Ethyl-2-pyridyl)ethoxy]aniline
19	646519-84-8	5-[[4-[2-(5-Ethyl-2-pyridinyl)-2-hydroxyethoxy]phenyl]methylene]-2,4-thiazolidinedione
20	101931-00-4	5-[[4-[2-(5-Ethyl-2-pyridinyl)-2-hydroxyethoxy]phenyl]methyl]-2,4-thiazolidinedione
21	646519-89-3	2,4-Thiazolidinedione, 5-[[4-[2-chloro-2-(5-ethyl-2-pyridinyl)ethoxy]phenyl]methyl] hydrochloride
22	96-33-3	Methyl acrylate
23	140-88-5	Ethyl acrylate
24	105355-25-7	Methyl 2-bromo-3-[4-[2-(5-ethyl-2-pyridyl)ethoxy]phenyl]propionate
25	62-56-6	Thiourea
26	105355-26-8	5-[4-[2-(5-Ethyl-2-pyridyl)ethoxy]benzyl]-2-imino-4-thiazolidinone
27	111025-46-8	Pioglitazone
28	112529-15-4	Pioglitazone hydrochloride

IV.1.2 MODELING ROUTES ADOPTING PROCESS SIMULATORS

After the selection of the twelve different designs described above, an equal number of flowsheets have been modelled in order to obtain mass and energy balances of each alternative (Figure IV-4). SuperPro Designer has been chosen as simulation environment for its simplicity and capability of dealing with batch processes. Indeed, more sophisticated process simulators need further unknown parameters (reaction kinetics, solubility of chemicals in various solvents, etc.) to perform a simulation of a batch system. Unfortunately, in most cases these supplementary information are not available in the literature and need to be estimated, requiring extra time and adding uncertainty to the sustainability evaluation. Therefore, for a preliminary analysis, data promptly available in literature have been adopted, i.e. setting a priori yields of reactions, temperature and pressure conditions, purity of extractions and residence times as they were specified in patents.

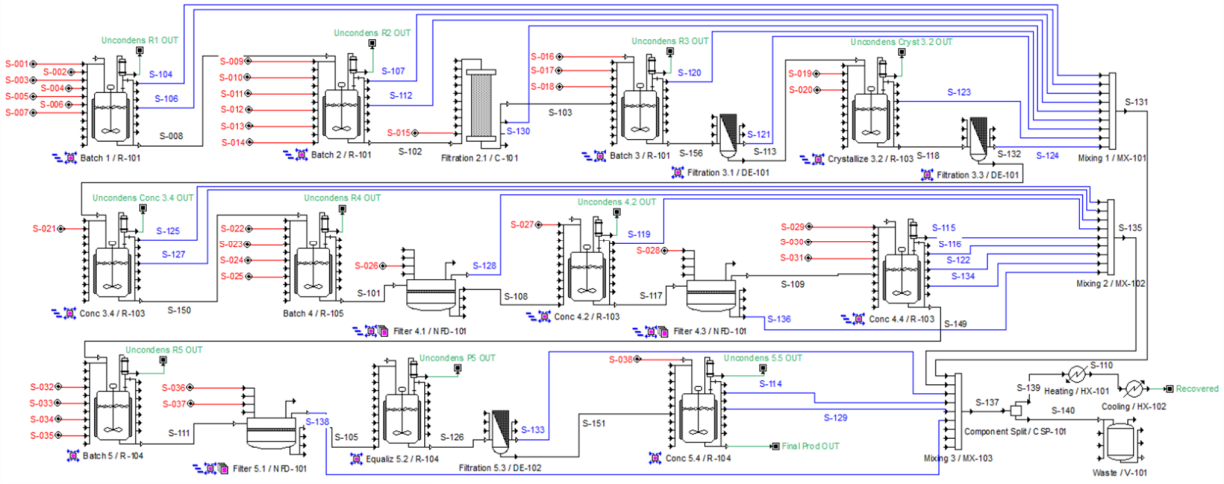

Figure IV-4. Model of Flowsheet 1.3 developed in SuperPro Designer

While the lack of literature data is well known for pharmaceutical compounds and their precursors, extensively studied chemicals as solvents don't require a large additional effort to be embedded within a detailed simulation in order to assess their recovery. This task has been carried out in Aspen Plus, developing a dedicated simulation for each different design and adopting WILSON-RK as thermodynamic method. The aim was to evaluate the energy consumption related to the recovery of solvents using traditional methods under the following assumptions:

- due to the variety of solvents adopted in each flowsheet, a minimum percentage of utilization has been set in order to assess which were the chemicals worth to recover. Thus, each detailed recovery simulation comprehends chemicals with a percentage higher than 5% to the total mass of substances involved in each process;
- the outlet streams of the each simulation have been investigated looking for the aforementioned chemicals in order to group them into different equalizing tanks containing the same compounds;
- the load of each tank containing one component (as said before the solutes involved less than 5% have been excluded from the detailed simulation) at normal condition has been evaporated and then brought back to normal conditions to consider the energy needed to reach the boiling point and the latent heat of vaporization in order to achieve a separation from the solutes;
- whether two or more component are mixed, a battery of distillation columns has been simulated depending on the number of components and the existence of azeotropes. Aiming to provide a standard approach to all separations while estimate the energy consumption at minimum reflux, we adopted RADFRAC unit operations in which every column had 30 stages, which were usually in excess. Azeotropic mixtures have been separated using the pressure swing technique in the range between 1 and 10 atm obtaining at least 99% of purity in outlet streams. A sensitivity analysis has

been performed on each column to pinpoint the minimum energy configuration under these constraints.

IV.1.3 CHOOSING AND CALCULATING INDICATORS

Previous step provided mass and energy balances essential for calculation of indicators for sustainability evaluation. Furthermore, some additional information need to be retrieved in order to be able to calculate each indicator chosen.

First, the cost of chemicals involved in each process have to be identified, as this information is compulsory for the calculation of the CP (Eq.1). A manufacturer can easily retrieve these data consulting its supplier or, if any, employing in-house production. In this work, the suppliers embedded in Scifinder platform have been analyzed, considering the average selling price of a substance unit among them ($/100g for catalysts, $/kg for reactants, $/l for solvents). This approach doesn't take into account lower costs due to the economy of scale derived from a stock purchase, while it preserves the price relationship within each class of compounds, ensuring a reliable evaluation among different designs in which compounds of the same category have been replaced. The recovery of unreacted reagents, supported catalysts recovered by filtration and solvents (k_i in Eq.2) have been set to 0, 0.99 and 0.99, respectively. Moreover, the project lifetime (τ) has been fixed to 15 years and the discount rate (r_d) to 7%. Wroth factors[107] adopted for a quick estimation of the auxiliary costs of the equipment involved in the designs under study have been reported in Table IV-3.

Table IV-3. Wroth factor adopted

Unit	Wroth Factor
Process tank	4.1
Storage tank	3.5
Dead-end filter	2
Nutsche filter	3
Heat exchanger	4.8
All other equipments	3.5

Moving forward, H-Phrases for each substance have to be retrieved in order to assign to chemicals specific hazard classes for the calculation of PCR and PEI. For the majority of compounds, H-Phrases can be found in MSDSs or Scifinder database, whereas for some chemicals, which are not in the market like byproducts or intermediates, H-Phrases have not been specified yet.

Moreover, at this stage of process design, some of the byproducts are still unknown due to the lack of literature data and dearth of experimental tests. Therefore two scenarios have been created, assigning the minimum (0) or the maximum (5) hazard class to unknown chemicals, representing the best and the worst case scenario, respectively. The assignation of each compound to a specific hazard class has been achieved using a Python script, which has been developed in order to speed up the selection process and guarantee the accuracy of the class assignation through an automated procedure.

Table IV-4 and Table IV-5 show, respectively, PCR and PEI original and normalized values for both scenarios, while CP is not included as it is independent from hazard classes. For each table, the column titled Δ reports the increasing percentages between the scores for each flowsheet using the two scenarios, while the column $\widehat{\Delta}$ shows the increasing percentages between the normalized scores for each flowsheet for the two scenarios.

Table IV-4. PCR scenarios

PCR							
		Original Scores (x_{ij})			**Normalized Scores ($\hat{x}_{ij}$)**		
j	**Route**	**Hazard Class 0**	**Hazard Class 5**	**Δ**	**Hazard Class 0**	**Hazard Class 5**	**$\widehat{\Delta}$**
1	Flowsheet 1	$2.16 \cdot 10^6$	$2.22 \cdot 10^6$	2.97%	0.020	0.020	0.00%
2	Flowsheet 1.1	$2.67 \cdot 10^6$	$2.73 \cdot 10^6$	2.42%	0.065	0.065	0.00%
3	Flowsheet 1.2	$6.16 \cdot 10^6$	$6.21 \cdot 10^6$	0.85%	0.373	0.373	0.00%
4	Flowsheet 1.3	$1.09 \cdot 10^7$	$1.10 \cdot 10^7$	0.40%	0.796	0.796	0.00%
5	Flowsheet 2	$1.96 \cdot 10^6$	$2.00 \cdot 10^6$	1.92%	0.000	0.000	0.00%
6	Flowsheet 3	$9.28 \cdot 10^6$	$9.38 \cdot 10^6$	0.99%	0.653	0.653	0.00%
7	Flowsheet 4	$4.00 \cdot 10^6$	$4.03 \cdot 10^6$	0.90%	0.180	0.180	0.00%
8	Flowsheet 4.1	$5.44 \cdot 10^6$	$5.52 \cdot 10^6$	1.36%	0.312	0.312	0.00%
9	Flowsheet 4.2	$4.54 \cdot 10^6$	$4.58 \cdot 10^6$	0.80%	0.229	0.229	0.00%
10	Flowsheet 5	$5.20 \cdot 10^6$	$5.35 \cdot 10^6$	2.86%	0.297	0.297	0.00%
11	Flowsheet 6	$5.51 \cdot 10^6$	$5.70 \cdot 10^6$	3.28%	0.327	0.327	0.00%
12	Flowsheet 7	$1.30 \cdot 10^7$	$1.33 \cdot 10^7$	2.54%	1.000	1.000	0.00%

Table IV-5. PEI scenarios

PEI							
		Original Scores (x_{ij})			**Normalized Scores ($\hat{x}_{ij}$)**		
j	**Route**	**Hazard Class 0**	**Hazard Class 5**	Δ	**Hazard Class 0**	**Hazard Class 5**	$\hat{\Delta}$
1	Flowsheet 1	$4.03 \cdot 10^5$	$4.60 \cdot 10^5$	12.20%	0.000	0.000	0.00%
2	Flowsheet 1.1	$4.46 \cdot 10^5$	$5.02 \cdot 10^5$	11.17%	0.009	0.009	0.00%
3	Flowsheet 1.2	$2.13 \cdot 10^6$	$2.17 \cdot 10^6$	2.06%	0.378	0.378	0.00%
4	Flowsheet 1.3	$3.44 \cdot 10^6$	$3.48 \cdot 10^6$	1.07%	0.666	0.666	0.00%
5	Flowsheet 2	$7.13 \cdot 10^5$	$7.45 \cdot 10^5$	4.37%	0.063	0.063	0.00%
6	Flowsheet 3	$2.90 \cdot 10^6$	$2.98 \cdot 10^6$	2.65%	0.557	0.557	0.00%
7	Flowsheet 4	$1.86 \cdot 10^6$	$1.89 \cdot 10^6$	1.64%	0.316	0.316	0.00%
8	Flowsheet 4.1	$2.36 \cdot 10^6$	$2.42 \cdot 10^6$	2.44%	0.433	0.433	0.00%
9	Flowsheet 4.2	$2.11 \cdot 10^6$	$2.14 \cdot 10^6$	1.46%	0.372	0.372	0.00%
10	Flowsheet 5	$1.87 \cdot 10^6$	$2.00 \cdot 10^6$	6.52%	0.340	0.340	0.00%
11	Flowsheet 6	$2.40 \cdot 10^6$	$2.53 \cdot 10^6$	5.25%	0.458	0.458	0.00%
12	Flowsheet 7	$4.70 \cdot 10^6$	$4.99 \cdot 10^6$	5.75%	1.000	1.000	0.00%

It is worth underlying how a variation in the original value of each indicator, due to the different hazard class assigned, is not affecting the normalized values (column $\hat{\Delta}$), as the normalization process, using Eq.46, keeps constant the relationships between the flowsheets scores.

$$\hat{x}_{ij} = \frac{x_{ij} - x_{i,min}}{x_{i,max} - x_{i,min}} \qquad Eq.46$$

where $\hat{x}_{ij}$ denotes the normalized value of the original indicator score, x_{ij}, for ith scenario and the jth flowsheet, in the range defined by the maximum, $x_{i,max}$, and the minimum, $x_{i,min}$, scores for the indicator under study.

Since normalized scores for DEA have been used, the identification of the best performing DMU will be independent from the adoption of a specific scenario. On the other hand, the choice of a particular scenario will affect the retrofit analysis, since the byproducts contribution to the final indicators scores will vary, which may lead to a different selection of the most valuable parameters in order to improve the suboptimal designs. In this work, the worst case scenario has been adopted, assigning the maximum

hazard class (5) to every unknown compound, since the utilization of brominated and chlorinated compounds may lead to the undesired release of their halogenated content.

Table IV-6 displays the calculated indicators scores, while Figure IV-5 shows the normalized scores for each indicator in a radar plot, in which the reader can infer that none of the flowsheets performs better than the others simultaneously considering all the indicators of interest, due to the existence of intersections between the lines connecting the values attained by each flowsheet for each indicator.

Table IV-6. Cost of Project (CP), Potential Chemical Risk (PCR) and Potential Environmental Impact (PEI) original and normalized scores

		Original			**Normalized**		
		Economics	**Social**	**Environmental**	**Economics**	**Social**	**Environmental**
#	**Routes**	**CP**	**PCR**	**PEI**	**CP norm**	**PCR norm**	**PEI norm**
1	Flowsheet 1	$4.21 \cdot 10^{8}$	$2.22 \cdot 10^{6}$	$4.60 \cdot 10^{5}$	0.501	0.020	0.000
2	Flowsheet 1.1	$5.79 \cdot 10^{8}$	$2.73 \cdot 10^{6}$	$5.02 \cdot 10^{5}$	0.931	0.065	0.009
3	Flowsheet 1.2	$3.39 \cdot 10^{8}$	$6.21 \cdot 10^{6}$	$2.17 \cdot 10^{6}$	0.277	0.373	0.378
4	Flowsheet 1.3	$4.53 \cdot 10^{8}$	$1.10 \cdot 10^{7}$	$3.48 \cdot 10^{6}$	0.588	0.796	0.666
5	Flowsheet 2	$2.37 \cdot 10^{8}$	$2.00 \cdot 10^{6}$	$7.45 \cdot 10^{5}$	0.000	0.000	0.063
6	Flowsheet 3	$5.78 \cdot 10^{8}$	$9.38 \cdot 10^{6}$	$2.98 \cdot 10^{6}$	0.931	0.653	0.557
7	Flowsheet 4	$4.07 \cdot 10^{8}$	$4.03 \cdot 10^{6}$	$1.89 \cdot 10^{6}$	0.463	0.180	0.316
8	Flowsheet 4.1	$5.53 \cdot 10^{8}$	$5.52 \cdot 10^{6}$	$2.42 \cdot 10^{6}$	0.860	0.312	0.433
9	Flowsheet 4.2	$4.19 \cdot 10^{8}$	$4.58 \cdot 10^{6}$	$2.14 \cdot 10^{6}$	0.495	0.229	0.372
10	Flowsheet 5	$6.04 \cdot 10^{8}$	$5.35 \cdot 10^{6}$	$2.00 \cdot 10^{6}$	1.000	0.297	0.340
11	Flowsheet 6	$5.30 \cdot 10^{8}$	$5.70 \cdot 10^{6}$	$2.53 \cdot 10^{6}$	0.798	0.327	0.458
12	Flowsheet 7	$5.96 \cdot 10^{8}$	$1.33 \cdot 10^{7}$	$4.99 \cdot 10^{6}$	0.978	1.000	1.000

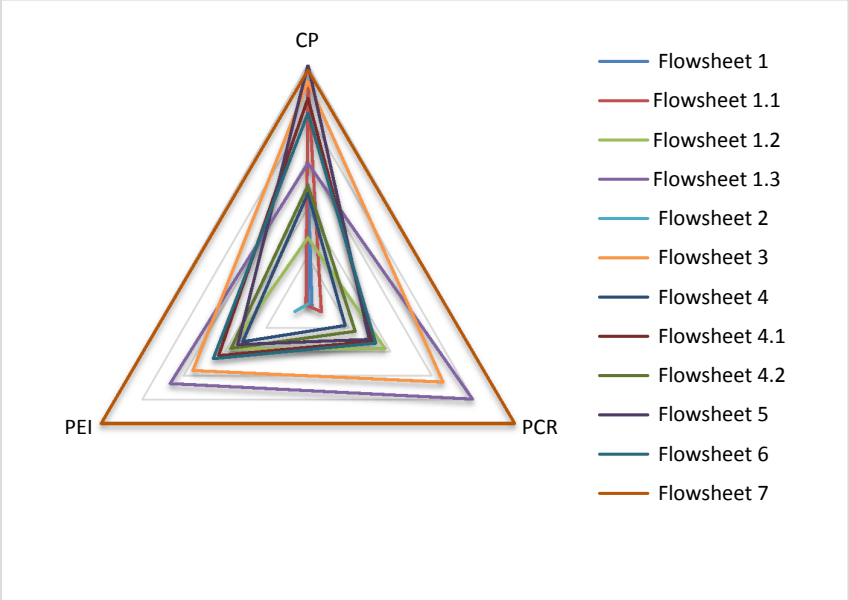

Figure IV-5. CP, PCR and PEI scores of the twelve flowsheets under study

The interpretation of Figure IV-5 can generate confusion due to a multitude of overlapping lines representing the flowsheets under investigation. Therefore, a deeper analysis using DEA becomes necessary in order to identify the optimal process design.

IV.1.4 Applying Data Envelopment Analysis (DEA) on indicators scores

Aiming to select the most efficient design, DEA has been performed, identifying each flowsheet as a DMU and considering CP, PCR and PEI, as, respectively, economic, social and environmental inputs criteria to be minimized in the analysis. To prevent the generation of numerical problems in the DEA models,[236] the indicators scores have been normalized using the aforementioned Eq.46.

The VRS DEA model features 17 variables and 5 constraints. It was implemented in GAMS 24.7.4 and solved with CPLEX 12.6.3.0 on an Intel® Core™ i5-480M processor operating at 2.66 GHz.[237,238] It took around 0.015 CPU seconds to solve every instance to global optimality.

The solution of the primal problem highlighted the existence of two efficient designs out of the twelve implemented as displayed in Figure IV-6. Inefficient designs show low efficiency scores (lower than 0.5), meaning that there is a considerable difference between efficient and inefficient alternatives, which are rather far from the efficient frontier.

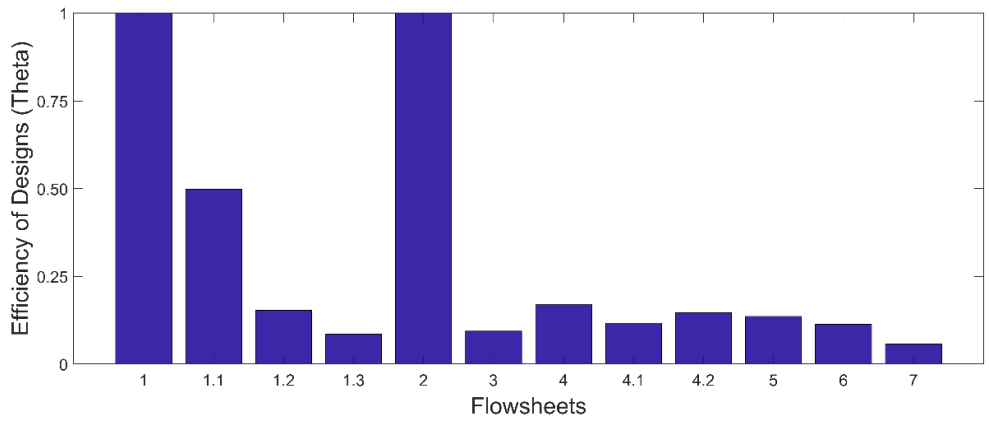

Figure IV-6. Relative efficiency of the 12 Flowsheets

In addition, the dual problem has been solved in order to identify the percentage improvements required for the inefficient design to move to the pareto-optimal frontier and become efficient using Eq.47.

$$\%\ improvement = \frac{original\ score - target\ score}{original\ score} \cdot 100 \qquad Eq.47$$

where target score is calculated using a linear combination of the efficient DMUs.

The percentage improvements for suboptimal designs are displayed using a heat map in Figure IV-7, in which the rows represent the inefficient synthesis routes and the columns embody the sustainability pillars quantified by indicators. The intensity of the yellow/red colors reveals the magnitude of the improvement needed for a design to become efficient, i.e. a darker color reflects a stronger target thus a bigger improvement effort.

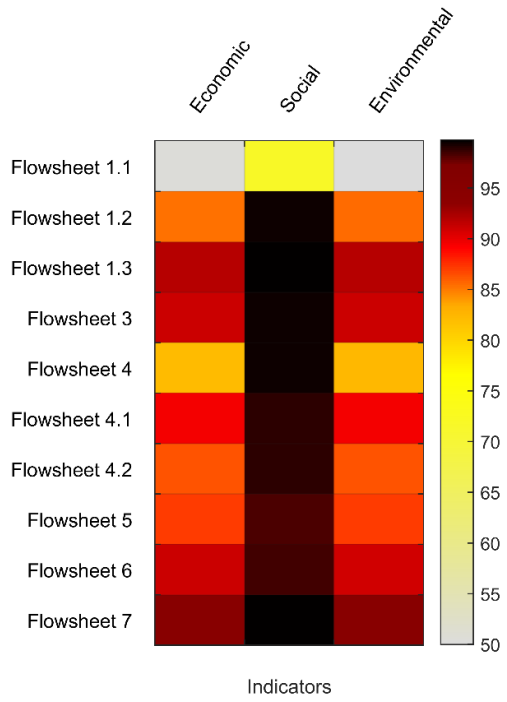

Figure IV-7: Inefficient improvements percentage

In this case study, the pareto optimal frontier depends on the linear combination of two efficient designs, whose indicators values approached zero rather closely. For this reason, the improvement targets of the inefficient units that seek to approach the Pareto frontier are very low and the percentage improvement required quite high. It is clear from Eq.47 how the improvement percentage increases as the target score decreases, bringing the overall ratio towards unity. Being the percentage of improvement assessed, the contribution of the various impacts on indicators scores need to be identified in order to evaluate the parameters that affect most the final indicators values. This will be performed in the following step of the methodology, while we still need to pinpoint the most sustainable design among the optimal ones.

The accomplishment of this the second task requires to perform again DEA using a super-efficiency analysis, i.e. a dual problem in which one efficient unit is removed from the model once at a time. This analysis follows the principles of the leave-one-out approach, evaluating how the efficient frontier modifies, as one efficient DMU is excluded from the system. However, this case study comprehends only two efficient designs, therefore a super-efficiency analysis would be meaningless, as the exclusion of one of them would remove the entire pareto frontier. Hence, the two efficient designs selected are the ones worth to be further investigated via experimental procedures, as they exhibited the minor impacts on a sustainability perspective.

IV.1.5 IMPROVING SUB-OPTIMAL DESIGNS

Sub-optimal process designs have been analyzed in order to identify the main sources of impacts, providing an insight of the variables that could contribute substantially to the improvement of indicators scores. Each indicator has been investigated separately, therefore we isolated the specific contribution to each sustainability aspect and identified possible common trends to every process design.

Contributions to economic impact have been reported in

Table IV-7, considering the costs evaluated using Eq.1 and Eq.2 for calculation of the CP. CapEx weights an average of 19.83% ± 6% on the overall project cost, meaning that Operative Expenditures are the main source of economic impact, although a reduction of CapEx, e.g. using in-house equipment, obtaining lower purchasing prices, reducing equipment sizes, etc., is still profitable. Within OpEx, the main source of expenses originates from raw materials cost (87.62% ± 6%), followed by LabEx (10.68% ± 6%).

Table IV-7. Relative percentage contributions to CP value

Flowsheet	***1.1***	***1.2***	***1.3***	***3***	***4***	***4.1***	***4.2***	***5***	***6***	***7***	***Average***
Materials	86.88%	85.85%	81.58%	86.55%	91.41%	92.30%	90.60%	89.91%	87.58%	83.51%	87.62%
Utilities	2.39%	1.48%	1.78%	1.02%	0.57%	0.38%	0.63%	2.35%	2.63%	3.77%	1.70%
LabEx	10.73%	12.67%	16.63%	12.43%	8.02%	7.32%	8.77%	7.74%	9.79%	12.72%	10.68%
OpEx	84.24%	74.26%	76.60%	78.05%	79.79%	82.86%	80.90%	81.59%	85.10%	78.32%	80.17%
CapEx	15.76%	25.74%	23.40%	21.95%	20.21%	17.14%	19.10%	18.41%	14.90%	21.68%	19.83%

Raw materials cost weights for about 70% of the total CP value, thus a recovery of the most prominent chemicals becomes essential in a reduction of costs viewpoint. The five main source of raw materials expenditures for each inefficient design are shown in Figure IV-8. Each contribution is related to the mass of substance involved in the process times the purchasing price, thus the reduction of the specific impact needs to be performed through the minimization of these entities. The recovery of the most utilized solvents and supported catalyst have already been set to 99%, considering the economic and environmental impacts related to their loss or emission, therefore our analysis underlines which other compounds revealed more valuable to be recovered, if possible. As expected, main reactants (2,4-Thiazolidinedione, 5-Ethyl-2-Pyridineethanol) arise as the biggest contribution to the final CP score, therefore maximum reaction yields are desirable.

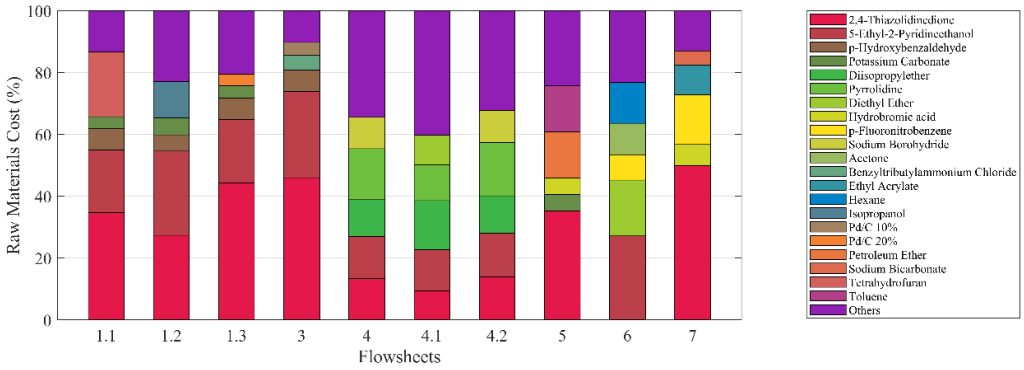

Figure IV-8. Five major contributions to raw materials cost for each inefficient design

The analysis of the contribution to the social impact follows an equivalent procedure. The PCR is calculated using Eq.7 that depends on a combination of mass flow and PCR hazard class related to each chemical, therefore Figure IV-9 shows the five most affecting chemicals for each flowsheet, reporting the contribution percentage of each one to the final PCR score for every inefficient process design. The adoption of hazardous solvents, e.g. methanol, chloroform, 1,2-dichloroethane, is clearly discouraged due to the detrimental combination of high mass flows and maximum hazard class. It is therefore beneficial to replace them with less harmful solvents belonging to lower hazard classes or to reduce the amount of chemical adopted per mass of final product.

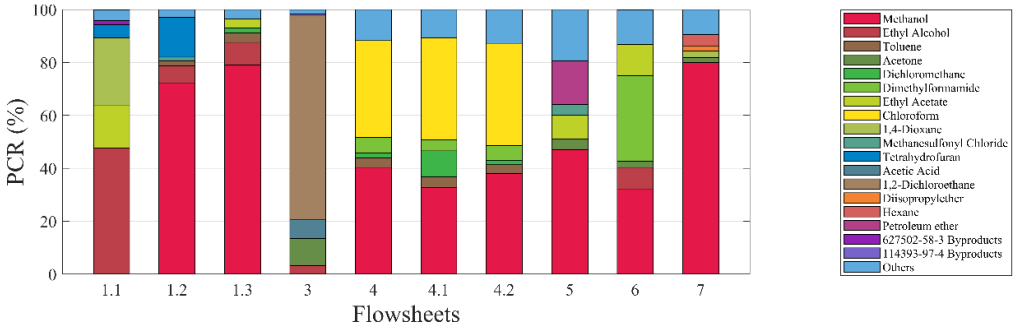

Figure IV-9. Five major chemical contributions to final PCR score for each inefficient design

Considering the environmental impact, the five chemicals that contributed most to the inefficient PEI final score have been identified for each process design, as shown in Figure IV-10. As previously mentioned, pollutant solvents, *e.g.* chloroform, 1,2-dichloroethane, 1,4-dioxane and methanol, are the main sources of environmental impact, since they are characterized by high massflow as well as high PEI hazard class. The byproducts generated during the production of intermediates of pioglitazone become more relevant than in PCR, since they have been assigned to the maximum PEI hazard class and their transfer coefficient (see Table II-4) has been set to the value 0.85 (liquid substance released in water streams).

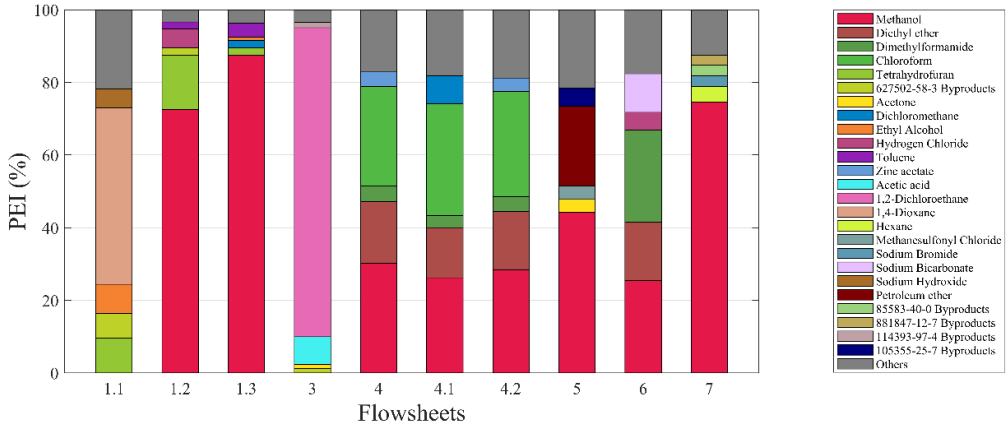

Figure IV-10. Major chemical contributions to final PEI score for each process design

Note that there might be many possible ways to make a suboptimal flowsheet efficient. This is mainly due to the existence of several types of projections, each leading to specific targets, as well as the possibility of controlling the emissions of chemicals in different ways. As an example, Flowsheet 1.1 could become efficient by reducing normalized values of indicators CP, PCR and PEI by 50, 71 and 50%, respectively. This could be accomplished for CP by recovering 72% of the outlet flow of four out of the five most influencing chemicals (considering that 5-Ethyl-2-Pyridineethanol is the main reactant and is totally consumed); for PCR and PEI by decreasing of 20.13% and 5.5%, respectively, the amount of the five most influencing chemicals in each indicator.

It is now necessary to deeper investigate the impact of process simulators on the accuracy of the results achieved. Several compounds involved in the synthesis routes were missing in the databases embedded within the simulators, therefore some assumptions have been taken in order to ensure the consistency of the results. Apart from the data available in patents (yields of reactions, temperature and pressure conditions, purity of extractions, residence time), a wide literature research has been

performed, comprehending Scifinder, Pubchem, MSDS, Chemspider, Chemical Book and Dortmund Database. For the parameters that were still missing, some predicted data provided by Scifinder and Chemspider have been adopted, using the same methodology within the same class of compounds, aiming to affect as little as possible the final accuracy. However, most of the properties have been regressed from experimental tests published in patents, assuming a high reliability of the data provided. Essentially, the process simulator reproduced the results obtained in a laboratory scale, increasing its production to an industrial size. The operating conditions have not been modified, with the purpose of avoiding the propagation of the regression error to unknown process conditions. Since the application of this methodology is restricted to a preliminary analysis of process designs, the results achieved through process simulations need to be confirmed via experimental tests. Once the parameters of the simulations have been confirmed, it is possible to generate novel process designs or alternative operating conditions in order to optimize the sustainability performance of the production.

Furthermore, it is possible to analyse the retrofit analysis results, aiming to infer the impact of estimated thermophysical properties on the final indicators scores. Since the main contributions to the final values of the indicators have been identified, it is rather important to determine the weight of the uncertainty of the predicted properties on the most influencing parameters:

- *Cost of the Project (CP):* The most influencing parameters are reported in Table IV-7. CapEx are independent from chemical properties and weights an average 19.83% ± 6% on the overall project cost. Within OpEx, the main source of expenses originates from raw materials cost (87.62% ± 6%), followed by LabEx (10.68% ± 6%), which is independent from chemical properties. The utilities cost (which is mainly affected by solvent recovery, based on well-established databases) shows a minimum impact on the overall analysis. Therefore, even if the estimated vapor pressures or heats of vaporization of unavailable compounds are highly uncertain, their contributions to final indicators scores will be negligible. Raw materials cost weights for about 70% of the total CP value, thus it is essential to use the proper amount of chemicals within the chemical process. Ratio among chemicals employed in the published patents have not been modified, hence the introduction of a source of uncertainty has been avoided. The prices are subjected to the market fluctuation, thus uncertainty is intrinsic in this parameter. In order to improve the accuracy, mean values from several chemical suppliers have been employed.
- *PCR and PEI*: As highlighted in Figure IV-9 and Figure IV-10, the major contributions to PCR and PEI final scores rise from the solvents employed in the processes. The ratio among solvents and reactants has been set using the data published in patents, therefore the main source of uncertainty does not reside on the process simulators.

IV.2 BIODIESEL FROM VEGETABLE OIL PRODUCTION PROCESS

Biodiesel production process consists on the conversion of vegetable oil or animal fats into a biofuel, through a transesterification and esterification reactions. The starting biomaterial reacts with short-chain alcohols, commonly methanol for the high conversion achieved or ethanol for its low purchasing price, catalyzed by either acids or bases, being the former less sensitive to both water and free fatty acids present in the oils, and the latter providing lower reaction times and purchasing cost.[239]

Vegetable oils are mainly composed of triglycerides, a family of compounds that belong to esters category, since they have been generated from the reactions of an acid (in this case three free fatty acids) with an alcohol (here glycerol, a trihydric alcohol). The transesterification process contemplates the deprotonation of the alcohol involved in the reaction with a strong base in order to exhibit its nucleophile peculiarity. The reaction kinetic is quite slow or even stationary under normal conditions, thus it is usually increased providing a source of heat as well as acid or basic catalysts (commonly potassium hydroxide, sodium hydroxide or sodium methoxide).

The majority of biodiesel production comes from virgin vegetable oils, characterized by low moisture (to avoid the undesirable hydrolysis of catalysts) and free fatty acids, and follows the synthesis route that employs the base-catalyzed technique as it is the most economical one, requiring only low temperatures and pressures and producing over 98% conversion yield. Since this is the predominant method for commercial-scale production, only the base-catalyzed transesterification process will be described below, presenting the three reactions in series that yield to the final product, which is a biodiesel based on a mixture of Fatty Acid Methyl Esters (FAME):

1) $C_3H_5(OCOR^{1,2,3})_3 + MeOH \leftrightarrow C_3H_5(OH)(OCOR^{2,3})_2 + MeCOOR^1$
tryglyceride methanol diglyceride methyl ester
2) $C_3H_5(OH)(OCOR^{2,3})_2 + MeOH \leftrightarrow C_3H_8(OH)_2(OCOR^3) + MeCOOR^2$
dyglyceride methanol monoglyceride methyl ester
3) $C_3H_8(OH)_2(OCOR^3) + MeOH \leftrightarrow C_3H_5(OH)_3 + MeCOOR^3$
monoglyceride methanol glycerol methyl ester

The mechanism underneath the transesterification reaction consists of a series of nucleophilic attacks by the alkoxide on the carbonyl carbons of the trygliceryde ($R^{1,2,3}$) starting material, obtaining a tetrahedral intermediate, which either reverts to the starting material, or proceeds to the transesterified product ($MeCOOR^{1,2,3}$) depending on the equilibrium reached by the overall system.

Using a literature base case, the investigation of different reactor parameters has been performed for the sake of identifying the most sustainable operating conditions, following the scheme shown in Figure IV-11, which comprehends the total variety of indicators adopted for two parallel assessments.

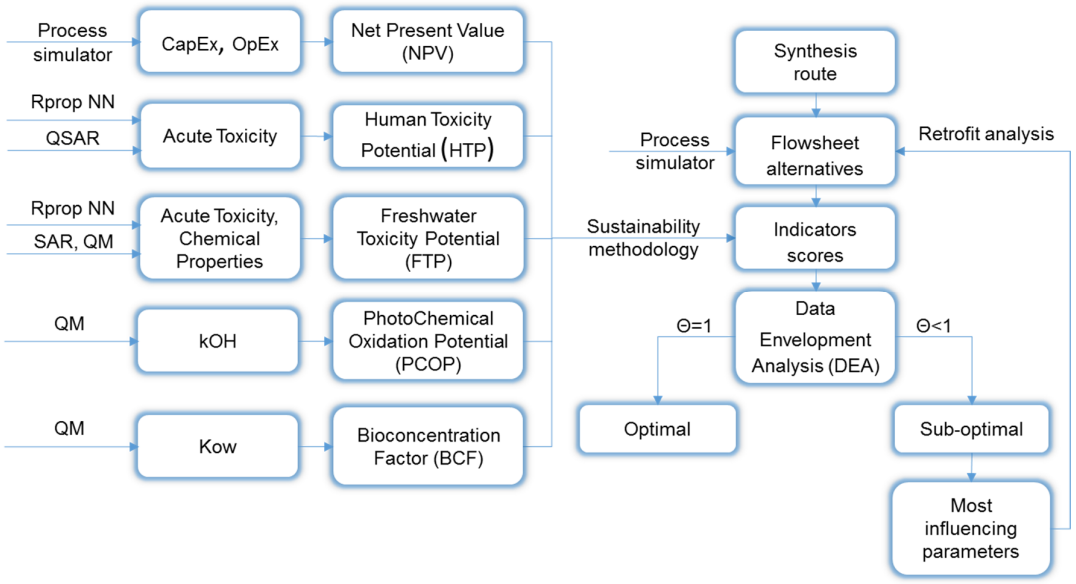

Figure IV-11. Decision path for sustainability assessment of biodiesel production process

IV.2.1 RETRIEVING DIFFERENT ROUTES

The different alternatives assessed in this work have been generated from the biodiesel production model proposed by Zhang,[240] which is a process for the alkali catalyzed production of biodiesel from palm oil, whose composition has been provided by Che Man.[241] Detailed descriptions of physical properties of starting materials and intermediates, as well as reliable transesterification kinetic parameters have been implemented within the model, enabling the practitioners to evaluate a wide number of alternatives through the modification of the operative conditions. Sodium hydroxide has been chosen as catalyst, which is neutralized by phosphoric acid addition to gain trisodium phosphate as precipitate salt. In this case study, a sensitivity analysis on the operative condition of the transesterification reactor has been performed. The parameters that have been evaluated are reported in Table IV-8, giving rise to 27 possible combinations.

Table IV-8. Reactor parameters evaluated in the sensitivity analysis

Temperature	***Pressure***	***Residence Time***
[°C]	***[bar]***	***[hours]***
50	3	1
60	4	2
70	5	3

IV.2.2 MODELING ROUTES ADOPTING PROCESS SIMULATORS

A reliable model has already been developed in a process simulator (Aspen Plus), including the overall biodiesel production, i.e. transesterification, methanol recovery, water washing, fatty acid methyl ester (FAME) purification, catalyst removal and glycerol purification, as shown in Figure IV-12.

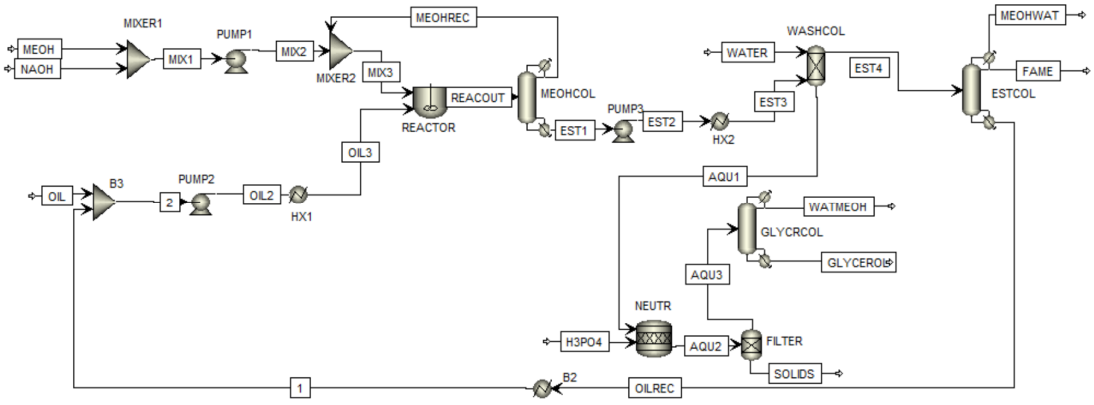

Figure IV-12. Biodiesel production process model developed in Aspen Plus

The base case reactor operative conditions were a temperature of 60 °C, a pressure of 4 bar and a residence time of 1 hour. Therefore, in order to evaluate the influence of each degree of freedom on the final product mass flow, a variance in both direction has been simulated for temperature and pressure, respectively of ±10 °C and ±1 bar, while an increasing of one and two hours has been implemented for the residence time. The setting of three values among three parameters generated 3^3=27 different alternatives to assess.

IV.2.3 CHOOSING AND CALCULATING INDICATORS

The various alternative models have been individually simulated and the results comprehending mass flows of each component involved in the process have been stored. The costs of each chemical have been retrieved in Scifinder,[154] while the value of palm oil and FAME biodiesel final product is based on average market prices retrieved online.[242,243] The capital costs and the utilities costs, required to calculate the operative ones, have been calculated using the add-in called Aspen Economics, integrated within Aspen Plus simulation. For NPV calculation (Eq.3), the project lifetime (τ) has been fixed to 15 years and the discount rate (r_d) to 7%, as for pioglitazone hydrochloride case study.

Social and environmental indicators have been evaluated using both PreADMET/COSMOtherm (Approach A) and a combination of Pallas, COSMOtherm and USETox techniques (Approach B). The former provides a quick evaluation of acute toxicity for human and freshwater, even though its reliability

is still to be established since it has not been widely used in literature yet. The latter employs the utilization of various software (some of which require a license) meaning that a good expertise on the utilization of different platforms is mandatory.

IV.2.3.1 ESTIMATION OF INDICATORS EMPLOYING PREADMET AND COSMO-RS (APPROACH A)

The techniques involved in the calculation of social and environmental indicators using the Approach A is reported in the scheme of Figure IV-13, where Aspen Plus has been employed as process simulator, PreADMET is the web tool based on resilient backpropagation neural network (Rprop NN) and COSMOtherm is the software elected for quantum mechanics (QM) calculation.

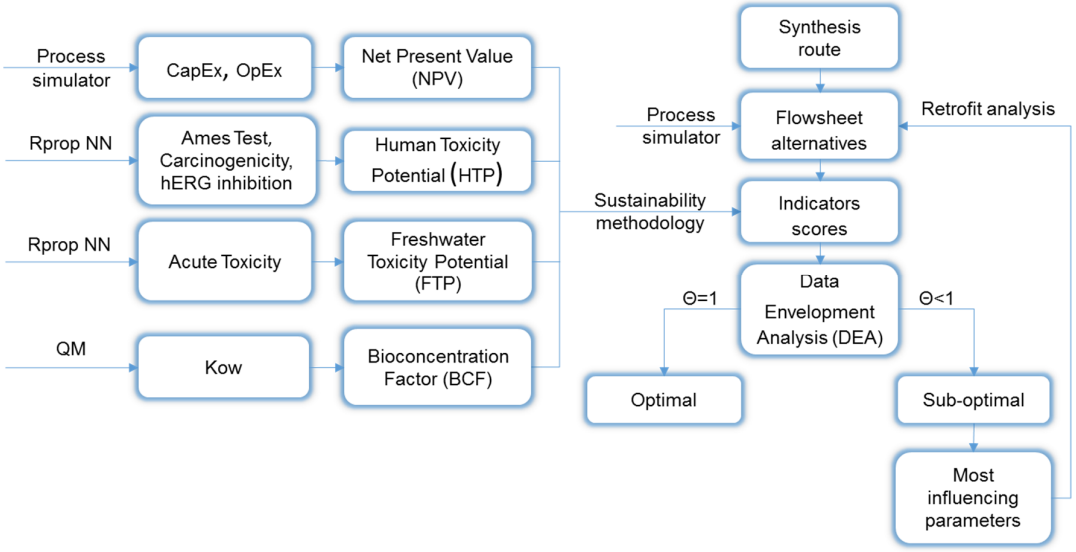

Figure IV-13. Flowchart describing the decision path for the sustainability assessment of biodiesel production process following Approach A.

PreADMET web tool has been used to in silico predict the results of Ames test, rodents' carcinogenicity and hERG inhibition in order to define the hazard class for HTP estimation. Furthermore, it has been used to forecast the values of acute concentration threshold for different aquatic organisms (algae, daphnia magna, feathed minnow and medaka) for the sake of calculate the FTP values, as described in Paragraph II.3.3.2. The 2D molecular structure was the only information required to run the estimation, since the calculation are based on artificial neural network algorithms.

In order to evaluate the performances of the methodology when more than three indicators are involved in the assessment, a fourth indicator, the Bioconcentration Factor (BCF), has been calculated

using the procedure of Paragraph II.3.3.3 and COSMOtherm to estimate Kow values, albeit other software for Kow estimation are available free of charge (see Paragraph III.2). The σ-surfaces of a selection of chemicals employed in the process are shown in Figure IV-14, which have been employed to estimate the Kow values of Eq.15-18 for BCF calculation.

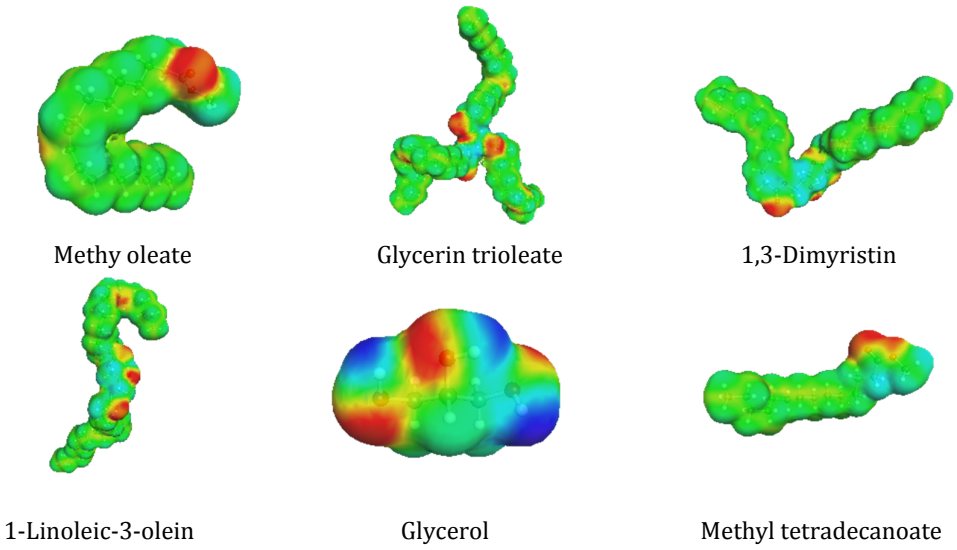

Figure IV-14. σ-surfaces obtained through the utilization of COSMO-RS for various molecules involved in the process. The blue color indicates a positive charge increase, surrounded by green neutral regions that turn into red as soon as the concentration of negative charge rises.

The normalized results of indicators forecast, using Eq.43, described so far are reported in Table IV-9.

Table IV-9. Net Present Value (NPV), Hazard Toxicity Potential (HTP), Freshwater Toxicity Potential (FTP) and BioConcentration Factor (BCF) normalized scores for the 27 operating conditions alternatives

	Operative conditions			Normalized Indicators			
Case	**Temp.**	**Press.**	**Residence Time**	**Economics**	**Social**	**Environmental**	**Environmental**
#	**[°C]**	**[bar]**	**[hours]**	**NPV**	**HTP**	**FTP**	**BCF**
1	50	3	1	0.8115	1	1	0.8726
2	50	3	2	0.9998	0.8985	0.8200	0.8560
3	50	3	3	0.0697	0.9051	0.7770	0.8554
4	50	4	1	0.8086	0.9715	0.9770	0.8728
5	50	4	2	1	0.8886	0.8110	0.8565
6	50	4	3	0.0743	0.8990	0.7713	0.8552
7	50	5	1	0.8088	0.9837	0.9880	0.8732
8	50	5	2	0.9973	0.9100	0.8314	0.8569
9	50	5	3	0.0661	0.9185	0.7897	0.8556
10	60	3	1	0.6171	0.7958	0.8154	0.9289
11	60	3	2	0.7756	0.8022	0.7660	0.9200
12	60	3	3	$3.21 \cdot 10^{-4}$	0	0	0
13	60	4	1	0.6169	0.8393	0.8551	0.9334
14	60	4	2	0.7735	0.8022	0.7660	0.9200
15	60	4	3	$1.604 \cdot 10^{-4}$	$3.349 \cdot 10^{-6}$	$5.030 \cdot 10^{-6}$	$5.410 \cdot 10^{-5}$
16	60	5	1	0.6168	0.8393	0.8551	0.9334
17	60	5	2	0.7733	0.8022	0.7660	0.9200
18	60	5	3	0	$8.826 \cdot 10^{-5}$	$8.196 \cdot 10^{-5}$	$1.126 \cdot 10^{-4}$
19	70	3	1	0.5071	0.8285	0.8557	0.9994
20	70	3	2	0.7306	0.8061	0.8085	0.9885
21	70	3	3	0.8796	0.8032	0.7940	0.9849
22	70	4	1	0.5086	0.8648	0.8921	1
23	70	4	2	0.7321	0.8470	0.8472	0.9901
24	70	4	3	0.8810	0.8034	0.7942	0.9849
25	70	5	1	0.5100	0.8651	0.8923	1.0000
26	70	5	2	0.7511	0.8473	0.8475	0.9901
27	70	5	3	0.8800	0.8037	0.7944	0.9849

The same indicators values are shown graphically in Figure IV-15, with the purpose of supporting the identification of the flowsheets that minimize the overall impact considering the indicators simultaneously, thus the radar plot shows the scores attained by every flowsheet for each one of the four indicator. The sensitivity analysis induced a modification on the amount of substances in specific streams (FAME, SOLIDS, GLYCEROL), while the nature of the implemented chemicals has been preserved.

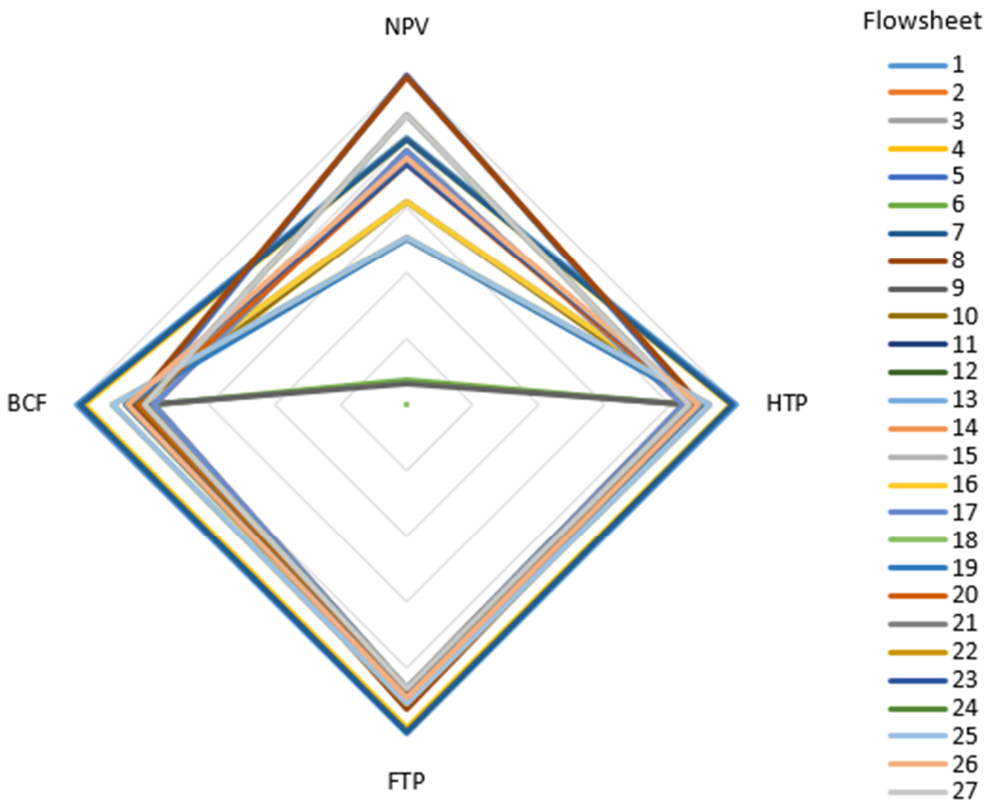

Figure IV-15. NPV, HTP, FTP and BCF scores of the 27 flowsheets under study

Three alternatives (Flowsheets 12, 15 and 18) perform efficiently, since they are overlapping on the central dot in the radar plot, although the choice of the best one is not straightforward. Therefore, the adoption of DEA is compulsory in order to confirm the hypothesis inferred from the plot and to detect the most efficient operating conditions.

IV.2.3.2 Estimation of indicators employing Pallas, COSMO-RS and USEtox (Approach B)

The calculation of HTP and FTP using Approach B follows the scheme presented in Figure IV-16, in which Aspen Plus is used as process simulator, Pallas is the software appointed for HTP QSAR acute toxicity forecast (see Paragraph II.3.2.2), ECOSAR (based on SAR) is the source of LC_{50} estimated values for fish, crustacean and algae, and COSMOtherm (based on QM) provides the chemical properties estimations, i.e. vapor pressure, Henry's Law constant, solubility in water at 25°C, Kow, and Koc, implemented within USEtox excel calculation sheet for FTP evaluation (Paragraph II.3.3.2). Furthermore, COSMO-RS is the quantum mechanical model adopted for the estimation of kOH, i.e. the reaction rate with hydroxyl radicals, employed in the calculation of PCOP (Paragraph II.3.3.4).

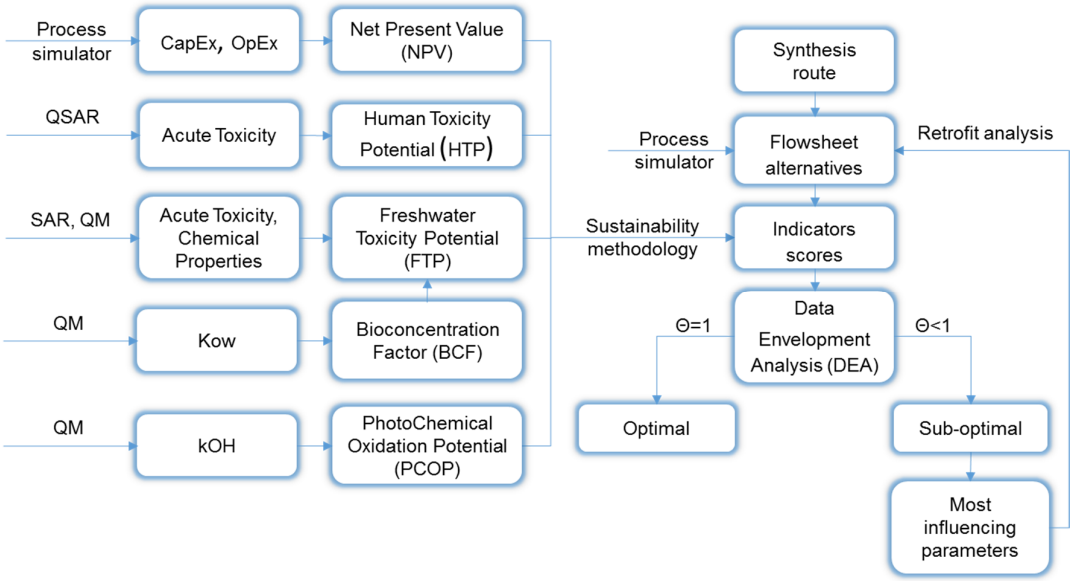

Figure IV-16. Flowchart describing the decision path for the sustainability assessment of biodiesel production process following Approach B

The sustainability assessment performed using Approach B embeds the BCF values within the chemical properties specified during the definition of substances database of USEtox. Unlike Approach A, in this analysis a novel contribution to environmental impact has been considered, i.e. the Photochemical Oxidation Potential, expanding the analysis perspective to miscellaneous natural elements. The normalized scores obtained for each one of the four indicators have been reported in the following Table IV-10.

Table IV-10. Net Present Value (NPV), Hazard Toxicity Potential (HTP), Freshwater Toxicity Potential (FTP) and PhotoChemical Oxidation Potential (PCOP) normalized scores for the 27 operating conditions alternatives

	Operative conditions			Normalized Indicators			
Case	**Temp.**	**Press.**	**Residence Time**	**Economics**	**Social**	**Environmental**	**Environmental**
#	**[°C]**	**[bar]**	**[hours]**	**NPV**	**HTP**	**FTP**	**PCOP**
1	50	3	1	0.8115	1	1	1
2	50	3	2	0.9998	0.9157	0.8200	0.9304
3	50	3	3	0.0697	0.9032	0.7770	0.9371
4	50	4	1	0.8086	0.9848	0.9770	0.9785
5	50	4	2	1	0.9098	0.8110	0.9216
6	50	4	3	0.0743	0.8996	0.7713	0.9319
7	50	5	1	0.8088	0.9919	0.9880	0.9885
8	50	5	2	0.9973	0.9228	0.8314	0.9406
9	50	5	3	0.0661	0.9112	0.7896	0.9489
10	60	3	1	0.6171	0.8734	0.8154	0.8037
11	60	3	2	0.7756	0.8610	0.7660	0.8185
12	60	3	3	$3.208 \cdot 10^{-4}$	0	0	0
13	60	4	1	0.6169	0.8996	0.8550	0.8406
14	60	4	2	0.7735	0.8610	0.7660	0.8185
15	60	4	3	$1.604 \cdot 10^{-4}$	$3.103 \cdot 10^{-5}$	$5.026 \cdot 10^{-6}$	$1.368 \cdot 10^{-5}$
16	60	5	1	0.6168	0.8996	0.8551	0.8406
17	60	5	2	0.7733	0.8610	0.7660	0.8186
18	60	5	3	0	$1.051 \cdot 10^{-4}$	$8.196 \cdot 10^{-5}$	$9.321 \cdot 10^{-5}$
19	70	3	1	0.5071	0.8868	0.8557	0.7922
20	70	3	2	0.7306	0.8665	0.8085	0.7832
21	70	3	3	0.8796	0.8611	0.7939	0.7833
22	70	4	1	0.5086	0.9100	0.8921	0.8258
23	70	4	2	0.7321	0.8914	0.8472	0.8190
24	70	4	3	0.8810	0.8612	0.7942	0.7835
25	70	5	1	0.5100	0.9102	0.8923	0.8260
26	70	5	2	0.7511	0.8916	0.8475	0.8192
27	70	5	3	0.8800	0.8614	0.7944	0.7837

Despite the values of the indicators changed in comparison with Approach A, the qualitative behavior of the sustainability assessment using Approach B highlights that the best operating conditions are still the ones assigned to Flowsheets 12, 15 and 18, since they minimize the indicators values simultaneously as show in the radar plot of Figure IV-17, where the aforementioned flowsheets are overlapping on the dot in the center of the axes.

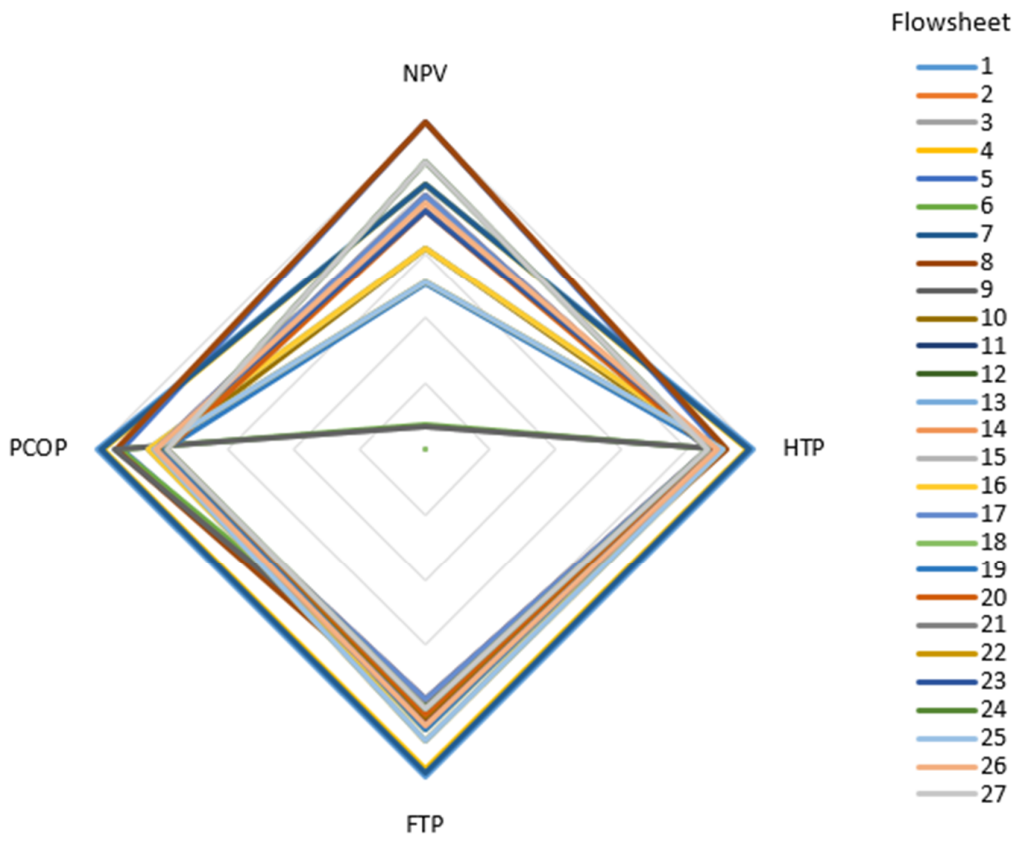

Figure IV-17. NPV, HTP, FTP and PCOP scores of the 27 flowsheets under study

As for Approach A, the identification of the best operating condition among the three that minimize the indicators value is difficult, since none of them performed better on each indicator simultaneously. Therefore, as suggested before, the adoption of DEA becomes essential in order to perform a more detailed screening of process alternatives.

IV.2.4 Applying Data Envelopment Analysis (DEA) on indicators scores

Since the results obtained from Approaches A and B are almost identical, with a negligible difference among the indicators values, DEA results have been reported for Approach B only, avoiding the repetition of indistinguishable figures. A first run of DEA algorithm has been performed with the purpose of gaining the values of efficiency (Θ) of every case study and thus certify that alternatives 12, 15 and 18 are the optimal ones indeed.

The VRS DEA model features 34 variables and 5 constraints. It was implemented in GAMS 24.7.4 and solved with CPLEX 12.6.3.0 on an Intel® Core™ i5-480M processor operating at 2.66 GHz.[237,238] It took around 0.015 CPU seconds to solve every instance to global optimality.

As it was expected, the solution of the primal problem highlights the existence of three efficient designs (number 12, 15 and 18 whose operating conditions have been reported in Table IV-11) out of the 27 implemented as displayed in
Figure IV-18. Inefficient designs show negligible efficiency scores (almost zero), meaning that there is a considerable difference between efficient and inefficient alternatives, which are rather far from the efficient frontier.

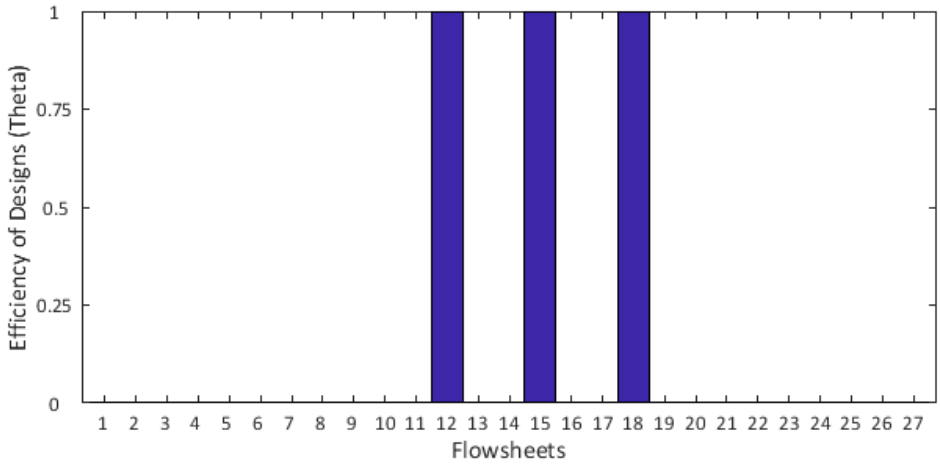

Figure IV-18. Relative efficiency of the 27 Flowsheets under investigation

Table IV-11. Operating conditions of the most efficient alternatives

Case	***Temperature***	***Pressure***	***Residence Time***
#	***[°C]***	***[bar]***	***[hour]***
12	60	3	3
15	60	4	3
18	60	5	3

Unlike the previous case study on pioglitazone hydrochloride, the improvement percentages of inefficient flowsheets calculated using the dual problem have not been reported, since the indicators scores are so low that all the improvement percentages are very close to 100%, generating a meaningless monochromatic heat map.

The identification of the most efficient design among the three optimal ones is still to be achieved, therefore DEA has been performed once again using a super-efficiency approach, i.e. a dual problem in which one efficient unit is removed from the model once at a time, following the leave-one-out principle. The results of the super-efficiency scores (Θ_{sup}) are shown in Table IV-12. Since the higher the value of Θ_{sup}, the more significant the influence of the process alternative on the Pareto frontier, the most efficient reactor operative conditions are the ones exhibited by case 15, i.e. reactor temperature of 60 °C, internal pressure of 4 bar and residence time of 3 hours.

Table IV-12. Super-efficiency scores for optimal designs

Case	Θ_{sup}
12	0.00005
15	1.781546
18	0.00016

IV.2.5 IMPROVING SUB-OPTIMAL DESIGNS

In this case study, the variables under investigation of the sustainability analysis have been the reactor operating conditions only, keeping constant the chemicals involved in the production process. Therefore, a different approach for the improvement of sub-optimal design has been carried on, operating on a qualitative level in order to comprehend the trend of the sustainability performance as soon as one of the degree of freedom (temperature, pressure or residence time) is modified.

Any temperature change from the base case value of 60 °C decreases the overall sustainability performances of the process, therefore a fixed value for this reactor parameter is recommended. This is justified since the three optimal alternatives have this fixed parameter in common.

The pressure maintained inside the reactor tank was set to 4 bar for the base case and, since the optimal alternatives exhibit a variation of this parameter, the influence of the pressure on the overall sustainability performances is less significant than the residence time and temperature.

The residence time of base case was fixed to one hour, while all the optimal flowsheets were characterized by a three hours residence time inside the reactor. This result allows practitioners to increase the transesterification yield prolonging the reaction time, since the growth of the operating cost due to a rise of the utility exploitation is counterbalanced by a better utilization of the raw materials.

IV.3 CdSe Quantum Dots Production

The Quantum Dots (QDs) materials class comprehend semiconductor nanoparticles that exhibit peculiar functional properties mainly related to their specific sizes. In fact, unique optical, electronic, magnetic and chemical properties revealed as soon as the crystal size approaches the Bohr radius (approximately the exciton size) specific for the designed material, due to quantum confinement effects.[244] Consequently, numerous applications of QDs emerged, encompassing various technology fields, e.g. biosensing and imaging,[245] photovoltaic,[246] photocatalysis,[247] and biology,[248] just to cite a few.

The fabrication of QDs developed through two main processes: the top-down approach, in which bulk material is crushed into nanosized particles,[249] and the bottom-up approach, which is based on chemical colloidal synthesis using building blocks.[249] Although the former provides a large quantity of QDs, the latter enables to achieve a better control on the nanoparticles size. Therefore, a significant effort has been spent on the optimization of synthesis conditions, since size uniformity of QDs is essential for a profitable application in optical and electronic devices.[250]

In this context, a series of laboratory tests on the bottom-up production of CdSe nanoparticles for photovoltaic application have been performed by Slejko,[251] aiming to maximize the yield of the process in prospect of the synthesis procedure scale-up to an industrial production.

The synthesis procedure is composed by the preparation of two separate solutions of nanocrystals precursors, which are then mixed up for the main reaction to occur in order to obtain the final product. The cadmium precursor solution has been prepared mixing cadmium oxide (CdO) and an excess of oleic acid (OA) dissolved in 1-octadecene (ODE) in order to obtain cadmium dioleate ($Cd(OA)_2$). The selenium precursor, trioctylphosphine selenide ($TOPSe$), is the product of the reaction between selenium powder and trioctylphosphine (TOP) in ODE. The combination of the two solutions starts the reaction between the precursors that is favored by the thermodynamic equilibrium, since the generation of nanoparticles removes the products from the solution, shifting the equilibrium to the right-hand side of the reaction.

$$CdO + 2OA \leftrightarrow Cd(OA)_2 + \frac{1}{2}O_2$$

$$Se + TOP \leftrightarrow TOPSe$$

$$Cd(OA)_2 + TOPSe \leftrightarrow CdSe + 2OA_{capping} + TOP_{capping}$$

Slejko's study focused on the kinetics of hot-injection synthesis of CdSe nanocrystals, exploring the system behavior as the growth temperature and the ratio between cadmium and its complexant (OA) have been systematically varied.

The sustainability evaluation methodology follows the scheme showed in Figure IV-19.

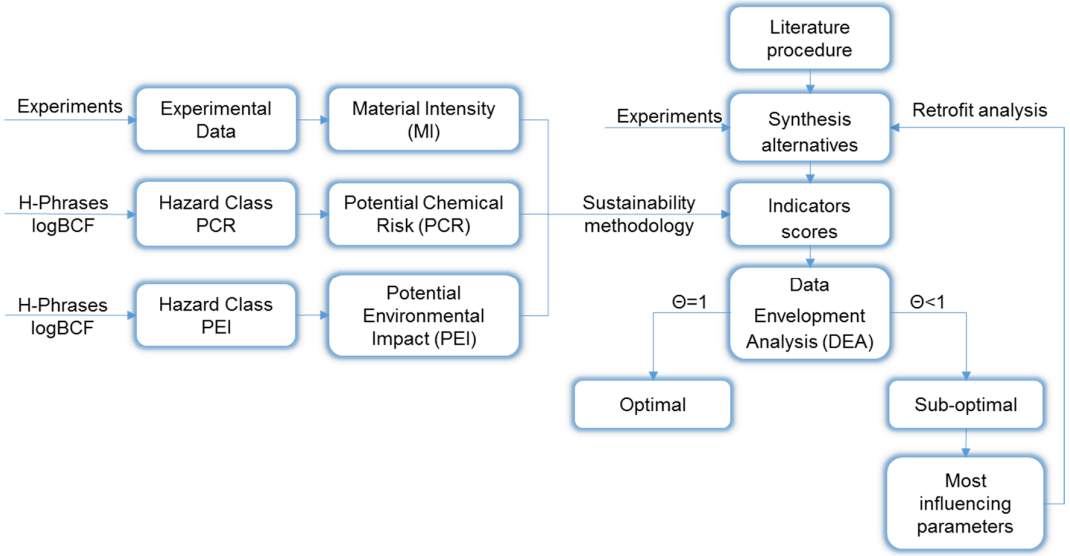

Figure IV-19. Decision path for sustainability assessment of CdSe quantum dots production process

IV.3.1 RETRIEVING DIFFERENT ROUTES

The alternative routes adopted for the sustainability assessment replicate the experimental synthesis performed by Slejko. The parameters that have been varied are the reaction temperature, for which the values of 210, 235, 255 and 270 °C have been chosen considering both the kinetic boost and the upper limit for ODE degradation, and the ratio between CdO and oleic acid, choosing 1:5, 1:6.5, 1:10, 1:15 and 1:19. The total combination of the various parameters values gave rise to 20 possible alternatives to choose from.

IV.3.2 MODELING ROUTES ADOPTING PROCESS SIMULATORS

The purpose of process modeling is providing the mass balance of the process and/or generating possible alternatives. Since this synthesis is still on a laboratory scale, the mass balances have already been supplied by experimental tests and the alternative reaction conditions have already been implemented in the set of experiments, there was no need to implement the synthesis process within a process simulator. However, as soon as the best operating conditions have been recognized and the reproducibility of the synthesis procedure has been guaranteed, the generation of the process model describing the procedure would be essential for scaling-up and tuning the operating parameters towards optimality through a sensitivity analysis coupled with an additional sustainability evaluation.

IV.3.3 CHOOSING AND CALCULATING INDICATORS

The sustainability indicators selected to assess the sustainability of the different alternatives have been the Profit Intensity (PI, see Paragraph II.3.1.3) for economic aspects, Potential Chemical Risk (PCR, see Paragraph II.3.2.1) for social impact and a combination of Potential Environmental Impact (PEI, see Paragraph II.3.3.1) and Bioconcentration Factor (BCF, see Paragraph II.3.3.3) for environmental issues.

The calculation of PI using Eq.6 required the retrieval of the raw materials costs, which have been purchased from Sigma Aldrich, the selling price of CdSe monodispersed nanoparticles and the total mass of saleable nanocrystals. The selling price of CdSe monodispersed nanoparticles has been calculated starting from the selling price of 100 €/mg for 6nm particle-size QDs. Since the particle-size distribution (PSD) is subjected to a normal curve, in which the wavelength of the first absorption peak depends on the synthesis temperature, the average size of the nanocrystals changed among the experiments. However, different-sized QDs exhibit peculiar properties related to their dimension, allowing a specific utilization for a wide range of dimensions: the main concern is then focused on the production of QDs as monodispersed as possible. In this perspective, the sealable fraction of monodispersed QDs among the total nanoparticles produced has been estimated using the full width at half maximum (FWHM) of the photoluminescence emission peak.

PCR has been calculated using Eq.7, and the H-Phrases required for its estimation have been retrieved on MSDSs.

Concerning PEI calculation, since the reaction occurs in liquid, the diffusion mean in which an environmental emission is more likely to happen is water (see Table II-4). Therefore BCF has been embedded within PEI calculation formula (Eq.48) in order to account for more environmental aspects using a single indicator:

$$PEI = \sum_i \frac{m_i^{max}}{m_{prod}} \cdot 10^{HCl_i^{PEI}} \cdot d_i \cdot log\, BCF_i \qquad Eq.48$$

The prediction of BCF is described in Paragraph II.3.3.3 and is based on Kow values estimated using COSMO-RS. Some illustrative examples of calculated σ-surfaces are displayed in Figure IV-20.

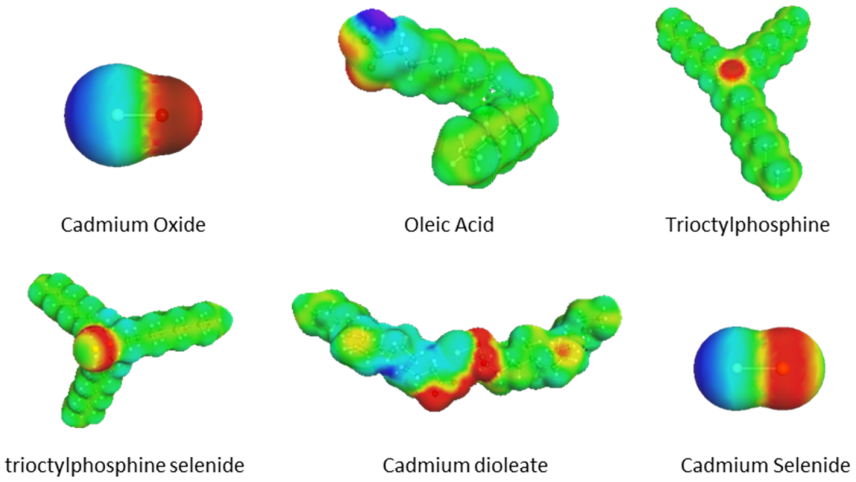

Figure IV-20. σ-surfaces calculated using COSMO-RS approach for various molecules involved in the process. The blue color indicates a positive charge increase, surrounded by green neutral regions that turn into red for negative charged areas

The contribution of the main product (CdSe QDs) on environmental and social impacts has been neglected since an increment on the yield of the process would have caused a worsening of the sustainability performances, invalidating the overall analysis.

Table IV-13 reports the normalized indicators scores relative to the experimental tests, that have been also represented using the radar plot of Figure IV-21.

Table IV-13. Profit Intensity (PI), Potential Chemical Risk (PCR) and the combination of Potential Environmental Impact (PEI) with BioConcentration Factor (BCF) normalized scores for the 20 experimental samples

	Experimental Parameters		***Normalized Indicators***		
Test	***Ratio***	***Temperature***	***Economics***	***Social***	***Environmental***
#	***[CdO]:[OA]***	***[°C]***	***PI***	***PCR***	***PEI:BCF***
1	1:5	210	0.2816	0.4893	0.4883
2	1:5	235	0.6413	0.5360	0.5352
3	1:5	255	0.8683	0.8090	0.8087
4	1:5	270	1	1	1
5	1:6.5	210	0.2653	0.6042	0.6038
6	1:6.5	235	0.5386	0.5991	0.5986
7	1:6.5	255	0.7948	0.7694	0.7693
8	1:6.5	270	0.6997	0.7281	0.7279

9	1:10	210	0	0.3854	0.3851
10	1:10	235	0.2665	0.3628	0.3625
11	1:10	255	0.3469	0.3739	0.3736
12	1:10	270	0.5346	0.5495	0.5495
13	1:15	210	0.1936	0.4381	0.4387
14	1:15	235	0.0775	0.1763	0.1762
15	1:15	255	0.2375	0.1925	0.1924
16	1:15	270	0.1914	0.1759	0.1758
17	1:19	210	0.4120	0.5090	0.5108
18	1:19	235	0.3837	0.3826	0.3847
19	1:19	255	0.1973	0	0
20	1:19	270	0.3547	0.3413	0.3432

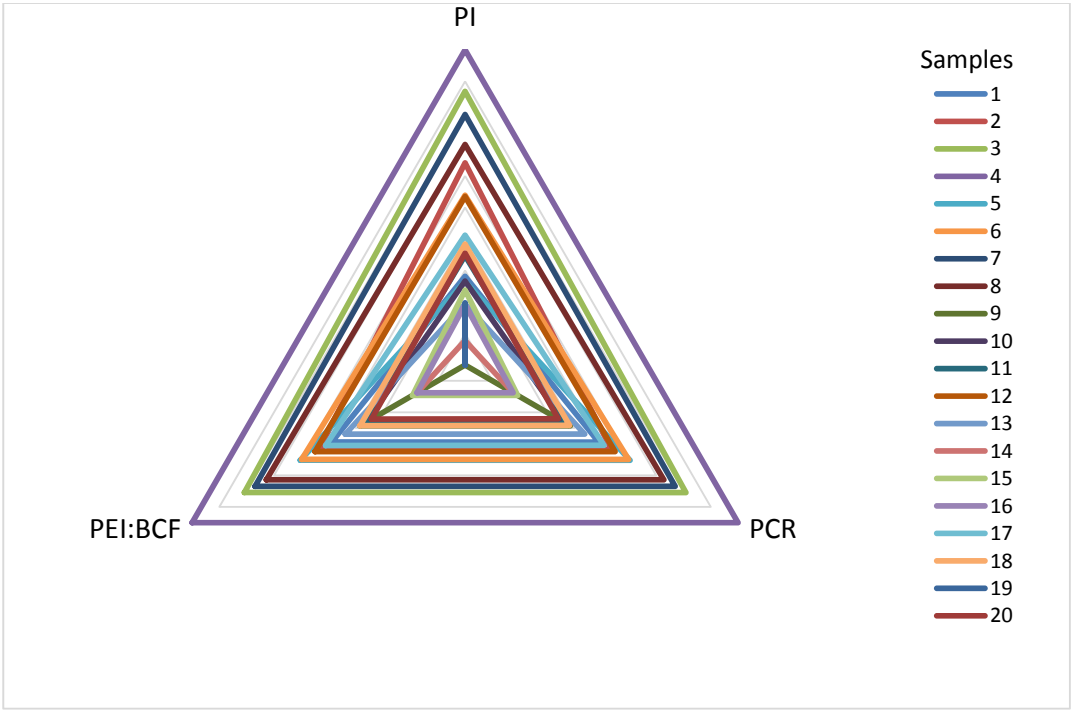

Figure IV-21. PI, PCR and PEI:BCF scores of the 20 experimental samples under study

Sample 4 gained the worst performance on every indicator simultaneously, emerging as the most unfavorable experimental test. Conversely, it is not clear which sample performed best, as the presence of intersections between samples towards the middle of the plot (where the scores are closer to optimality) requires a further analysis using DEA.

IV.3.4 APPLYING DATA ENVELOPMENT ANALYSIS

Initially, DEA has been run once in order to identify the efficient DMUs and thus obtain a first screening of the process designs. The VRS DEA model, featuring 25 variables and 5 constraints, was implemented in GAMS 24.7.4 and solved with CPLEX 12.6.3.0 on an Intel® Core™ i5-480M processor operating at 2.66 GHz,[237,238] requiring 0.015 CPU seconds to solve every instance to global optimality. The results of the relative efficiencies estimation using DEA is displayed in the column chart of Figure IV-22.

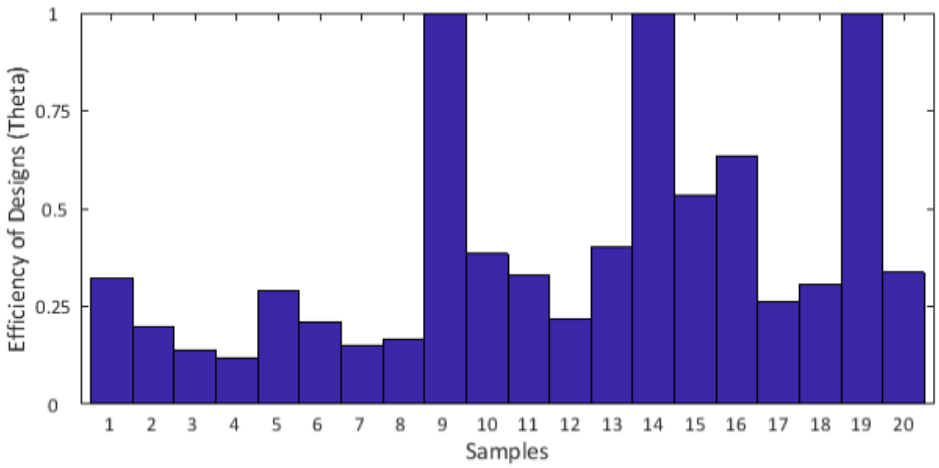

Figure IV-22. Results from the Dual VRS DEA showing the relative efficiencies of the 20 experimental samples

Three samples (9, 14, 19) out of 20 emerged as efficient, since their Θ values equal unity. Differently from biodiesel production case study, the optimal alternatives do not share any of the changeable parameters values, as reported in Table IV-14, where the results from a DEA Dual super-efficiency model has been additionally reported, aiming to select the most sustainable design among the efficient ones using the leave-one-out rule.

Table IV-14. Experimental settings and super-efficiency scores of the most efficient alternatives

Sample	***Ratio***	***Temperature***	$\boldsymbol{\Theta_{sup}}$
#	***[CdO]:[OA]***	***[°C]***	
9	1:10	210	0.07747
14	1:15	235	1.17632
19	1:19	255	0.35165

The most efficient parameters set up is the one exhibiting the highest Θ_{sup} value, therefore Sample 14, with a [CdO]:[OA] ratio of 1:15 and a growth temperature of 235 °C, performed best from a sustainability viewpoint.

IV.3.5 IMPROVING SUB-OPTIMAL DESIGN

The economic indicator is clearly strongly related to the amount of saleable nanocrystals produced, i.e. monodispersed ones, therefore any improvement towards this direction exerts a substantial leverage on the economic performance.

Since the nature of the substances involved is preserved among the various experimental tests, an analysis on the most affecting chemicals exhibited an analogous result for all the different alternatives: the most influencing chemical is ODE. This is due to the quantity of the solvent adopted, as the dilution of the solutes is extremely high. Therefore a decrease of the quantity of solvent involved in the process is suggested from an industrial application as well as on a sustainability viewpoint.

From the experimental data, Sample 9 is the one with the maximum number of monodispersed nanoparticles produced, while Sample 19 exhibits the best yield in terms of utilization of reactants, even if the nanocrystals are not as monodispersed as in Sample 9. The best alternative appears to be a trade-off between this two options, as Sample 14 exhibits a sharper dispersion of QDs than Sample 19 as well as a more efficient conversion of reactants to sealable nanocrystals than Sample 9.

CHAPTER V

CONCLUDING REMARKS

This dissertation has been focused on a new approach for the sustainability evaluation of chemical industrial processes, aiming to provide a fast preliminary screening of alternative process designs to be further evaluated using well-established assessment methodologies.

A multiscale approach has been adopted, using molecular simulation at the nanoscale as well as process simulation at the macroscale. Numerous illustrative software have been used, providing a wide set of possible tools to choose from.

An evaluation on the best techniques available for the Kow estimation has been performed considering several pharmaceutical compounds. Even though the numbers of chemicals involved in the analysis may be not sufficient to infer an exhaustive model, a preliminary selection of the one which is more in accordance with experimental values has been done, electing the QSAR ACD/Labs technique as the most accurate prediction method, nonetheless it is still in need of improvements. A deeper and extensive analysis comprehending a wider number of pharmaceutical compounds has been already scheduled for the next future.

The indicators selected for the sustainability evaluation, albeit they were not comprehensive of all the possible contributions, comprehended aspects belonging to each sustainability pillar, providing a general estimation of the process designs efficiency. Practitioners will experience all the benefits of this framework even though they choose to adopt any in-house or suitable indicators, since DEA will deal with inherent trade-offs. Indeed, the implementation of DEA within the overall selection process has given a mathematical fundamental to the efficiency definition of each process design, while it also provided the improvement targets in order to enhance the performances of suboptimal designs through a retrofit analysis.

This methodology benefits from a good versatility, confirmed by its applicability to several case studies belonging to various chemical industry fields, from pharmaceutical, to biochemical and nanotechnology. However, it is important to stress that the field of application for this methodology is limited to the battery limits of the production plant, omitting several impacts to sustainability, such as supply chain contributions or disposal of wastes, which need to be accounted for an exhaustive life cycle assessment.

Future perspectives include the introduction of uncertainty within the process simulations as well as the adoption of different software for the estimation of molecular properties in order to enhance the accuracy of the indicators employed.

ACKNOWLEDGEMENTS

First of all, I want to thank Bracco Corporate for the financial support provided for my project. Without their interest on my research topic, this dissertation would have never come to light.

Then I would like to express my gratitude to my supervisor, Professor Maurizio Fermeglia, who struggled during these three years in search of time to dedicate to my research, giving me precious indications towards fruitful directions.

I am sincerely thankful for the opportunity to collaborate with the research group of Dr. Gonzalo Guillen-Gozalbez, since our work together has been essential for the development of the methodology adopted in this work. I will always remember the months spent in London, thanks to Andres, Phantisa, Raul, Amjaad, Daniel, Michael and Yukon, who accepted me as a member of their awesome gang.

I feel that a special thank is mandatory to the one who shared every step of this PhD rollercoaster with me, Lele. We spent three years working and chilling together, supporting one another in every occasion and creating a real friendship from scratch.

I would like to acknowledge all the people who gave me suggestions and comments on the improvements achieved during my research project, in particular Massimo, Vanni, Paola, Erik and Sabrina. I wish our collaboration will be fruitful in the future as it has been during my PhD.

My biggest "thank you" goes to my parents, Carla e Roberto, whose love and encouragement have always been cornerstones of my life, providing me the necessary self-confidence to overcome any difficulty I had to face.

My friends and housemates have been essential for reaching a balance in my life outside of university, sharing with me hobbies and passion: for this reason, I thank them all.

Last but not least, I would like to thank my girlfriend Anna, to whom this dissertation is dedicated, for always been by my side, for her support during hard days and for the joy and love during the cheerful ones.

BIBLIOGRAPHY

(1) Scheffer, M.; Carpenter, S.; Foley, J. A.; Folke, C.; Walkerk, B. Catastrophic Shifts in Ecosystems. *Nature* **2001**, *413*, 591–596.

(2) Rockström, J.; Steffen, W.; Noone, K.; Persson, Å.; Chapin, F. S.; Lambin, E.; Lenton, T. M.; Scheffer, M.; Folke, C.; Schellnhuber, H. J.; *et al.* Planetary Boundaries: Exploring the Safe Operating Space for Humanity. *Ecol. Soc.* **2009**, *14*, 472–475.

(3) Steffen, W.; Richardson, K.; Rockström, J.; Cornell, S. E.; Fetzer, I.; Bennett, E. M.; Biggs, R.; Carpenter, S. R.; De Vries, W.; De Wit, C. A.; *et al.* Planetary Boundaries: Guiding Human Development on a Changing Planet. *Science (80-.).* **2015**, *347*.

(4) Vitousek, P. M.; Mooney, H. A.; Lubchenco, J.; Melillo, J. M. Human Domination of Earth' S Ecosystems. *Science (80-.).* **1997**, *277*, 494–499.

(5) Walther, G.-R.; Post, E.; Convey, P.; Menzel, A.; Parmesank, C.; Beebee, T. J. C.; Fromentin, J.-M.; Hoegh-Guldberg, O.; Bairlein, F. Ecological Responses to Recent Climate Change. *Nature* **2002**, *416*.

(6) Ramanathan, V.; Crutzen, P. J.; Kiehl, J. T.; Rosenfeld, D. Aerosols, Climate, and the Hydrological Cycle. *Science (80-.).* **2001**, *294*, 2119–2124.

(7) Foley, J. A.; Defries, R.; Asner, G. P.; Barford, C.; Bonan, G.; Carpenter, S. R.; Chapin, F. S.; Coe, M. T.; Daily, G. C.; Gibbs, H. K.; *et al.* Global Consequences of Land Use. *Science* **2005**, *309*, 570–574.

(8) Postel, S. L. Entering an Era of Water Scarcity : The Challenges Ahead. *Ecol. Appl.* **2000**, *10*, 941–948.

(9) Miller, R. G.; Sorrell, S. R.; Lane, H.; Kt, S.; Miller, R. G. The Future of Oil Supply. *Phil. Trans. R. Soc. A* **2013**, *372*.

(10) Anastas, P. T.; Zimmerman, J. B. Design through the 12 Principles of Green Engineering. *Environemental Sci. Technol.* **2003**, *37*, 94A–101A.

(11) Moss, R. H.; Edmonds, J. A.; Hibbard, K. A.; Manning, M. R.; Rose, S. K.; van Vuuren, D. P.; Carter, T. R.; Emori, S.; Kainuma, M.; Kram, T.; *et al.* The next Generation of Scenarios for Climate Change Research and Assessment. *Nature* **2010**, *463*, 747–756.

(12) Cambridge Dictionary https://dictionary.cambridge.org/ (accessed Jul 5, 2017).

(13) Präsidenten, B.; Projekte, E. Deep Roots – A Conceptual History of "Sustainable Development" (Nachhaltigkeit). **2007**.

(14) International Union for Conservation of Nature and Natural Resources. *World Conservation Strategy: Living Resource Conservation for Sustainable Development*; 1980.

(15) Brundtland Commission. *Our Common Future*; Oxford University Press: Oxford, 1987.

(16) Kates, Robert, W.; Parris, T. M.; Leiserowitz, A. A. What Is Sustainable Development? Goals, Indicators, Values, and Practice. *Environ. Sci. Policy Sustain. Dev.* **2005**, *47*, 8–21.

(17) Mebratu, D. Sustainability and Sustainable Development: Historical and Conceptual Review. *Environ. Impact Assess. Rev.* **1998**, *18*, 493–520.

(18) Johnston, P.; Everard, M.; Santillo, D.; Robèrt, K.-H. Reclaiming the Definition of Sustainability. *Environ. Sci. Pollut. Res. - Int.* **2007**, *14*, 60–66.

(19) Mosovsky, J.; Dickinson, D.; Morabito, J. Creating Competitive Advantage through Resource Productivity, Eco-Efficiency, and Sustainability in the Supply Chain. In *Proceedings of the 2000 IEEE International Symposium on Electronics and the Environment*; IEEE; pp. 230–237.

(20) Shaker, R. R. The Spatial Distribution of Development in Europe and Its Underlying Sustainability Correlations. *Appl. Geogr.* **2015**, *63*, 304–314.

(21) Bolis, I.; Morioka, S. N.; Sznelwar, L. I. When Sustainable Development Risks Losing Its Meaning. Delimiting the Concept with a Comprehensive Literature Review and a Conceptual Model. *J. Clean. Prod.* **2014**, *83*, 7–20.

(22) García-Serna, J.; Pérez-Barrigón, L.; Cocero, M. J. New Trends for Design towards Sustainability in Chemical Engineering: Green Engineering. *Chem. Eng. J.* **2007**, *133*, 7–30.

(23) Toma, L. Computer Aided Design of Sustainable Industrial Processes, Università di Padova, 2008.

(24) Jin, X.; High, K. A. A New Conceptual Hierarchy for Identifying Environmental Sustainability Metrics. *Environ. Prog.* **2004**, *23*, 291–301.

(25) Elkington, J. *Cannibals with Forks : The Triple Bottom Line of 21st Century Business*; Capstone, 1997.

(26) Geissdoerfer, M.; Savaget, P.; Bocken, N. M. P.; Hultink, E. J. The Circular Economy E A New Sustainability Paradigm? *J. Clean. Prod.* **2017**, *143*, 757–768.

(27) Investopedia http://www.investopedia.com/terms/e/economy.asp (accessed Jul 5, 2017).

(28) Merriam-Webster Dictionary https://www.merriam-webster.com/dictionary/environment (accessed Jul 5, 2017).

(29) Wikipedia https://en.wikipedia.org/wiki/Society (accessed Jul 5, 2017).

(30) Holmberg, J. *Policies for a Small Planet*; Earthscan Publications, 1994.

(31) United Nations Conference on Environment & Development. *Agenda 21*; Rio de Janerio, 1992.

(32) Grove, J. Triple Miracle Sees Huge Rise in EU Funds for Frontier Research. *Times High. Educ.* **2011**.

(33) European Commission. *Horizon 2020 – the EU's New Research and Innovation Programme*; Brussels, 2013.

(34) Reich-Weiser, C.; Dornfeld, D. A. A Discussion of Greenhouse Gas Emission Tradeoffs and Water Scarcity within the Supply Chain. *J. Manuf. Syst.* **2009**, *28*, 23–27.

(35) Cooper, R. G.; Edgett, S. J. Maximizing Productivity in Product Innovation. *Ind. Res. Institute, Inc* **2008**, *51*.

(36) Office of Technology Assessment (OTA). *Green Products by Design: Choices for a Cleaner Environment*; 1992.

(37) Fuller, D. A.; Ottman, J. A. Moderating Unintended Pollution: The Role of Sustainable Product

Design. *J. Bus. Res.* **2004**, *57*, 1231–1238.

(38) Singh, R. K.; Murty, H. R.; Gupta, S. K.; Dikshit, A. K. An Overview of Sustainability Assessment Methodologies. *Ecol. Indic.* **2009**, *9*, 189–212.

(39) Poveda, C. A.; Lipsett, M. A Review of Sustainability Assessment and Sustainability/Environmental Rating Systems and Credit Weighting Tools. *J. Sustain. Dev.* **2011**, *4*, 36.

(40) Sekulic, D. P. An Entropy Generation Metric for Non-Energy Systems Assessments. *Energy* **2009**, *34*, 587–592.

(41) Sikdar, S. K.; Sengupta, D.; Harten, P. More on Aggregating Multiple Indicators into a Single Index for Sustainability Analyses. *Clean Technol. Environ. Policy* **2012**, *14*, 765–773.

(42) Angelakoglou, K.; Gaidajis, G. A Review of Methods Contributing to the Assessment of the Environmental Sustainability of Industrial Systems. *J. Clean. Prod.* **2015**, *108*, 725–747.

(43) Shelton, J. D. An Investigation of Sustainability Metrics in Industry to Aid Product Design, Production, and Distribution Processes. **2010**.

(44) U.S. Environmental Protection Agency (EPA). *Safer Choice*; 2015.

(45) International Standard Organization. *ISO 14040 Series -Environmental Management – Life Cycle Assessment – Principles and Framework.*; Geneva, Switzerland.

(46) Bergeson, L. L.; Campbell, L. M. Economic Incentives for TQEM: Are They in Your Future? *Environ. Qual. Manag.* **1991**, *1*, 151–158.

(47) Srivastava, S. K. Green Supply-Chain Management: A State-of- the-Art Literature Review. *Int. J. Manag. Rev.* **2007**, *9*, 53–80.

(48) Ahi, P.; Searcy, C. A Comparative Literature Analysis of Definitions for Green and Sustainable Supply Chain Management. *J. Clean. Prod.* **2013**, *52*, 329–341.

(49) ISO International Standard Organization. ISO 14001 - Environmental Management https://www.iso.org/iso-14001-environmental-management.html (accessed Sep 15, 2017).

(50) Organisation for Economic Co-operation and Development (OECD). *Environmental Outlook for the Chemicals Industry*; 2001.

(51) Anastas, P. T.; Warner, J. C. *Green Chemistry: Theory and Practice*; Press, O. U., Ed.; Oxford, 1998.

(52) Anastas, P. T.; Heine, L.; Williamson, T. C. *Green Engineering*; Society, A. C., Ed.; 2000.

(53) European Union. *Regulation (EC) No 1907/2006 of the European Parliament and of the Council of 18 December 2006 Concerning the Registration, Evaluation, Authorisation and Restriction of Chemicals (REACH), Establishing a European Chemicals Agency*; 2006.

(54) Hoffman, A. J. Institutional Evolution and Change: Environmentalism and the U.S. Chemical Industry. *Acad. Manag. J.* **1999**, *42*, 351–371.

(55) Azapagic, A.; Perdan, S. Indicators of Sustainable Development for Industry: A General Framework. *Trans IChemE* **2000**, *78*, 243–261.

(56) Finnveden, G.; Hauschild, M. Z.; Ekvall, T.; Guinée, J.; Heijungs, R.; Hellweg, S.; Koehler, A.;

Pennington, D.; Suh, S. Recent Developments in Life Cycle Assessment. *J. Environ. Manage.* **2009**, *91*, 1–21.

(57) Tufvesson, L. M.; Tufvesson, P.; Woodley, J. M.; Börjesson, P. Life Cycle Assessment in Green Chemistry: Overview of Key Parameters and Methodological Concerns. *Int. J. Life Cycle Assess.* **2013**, *18*, 431–444.

(58) Thinkstep. GaBi Version: 8.0.

(59) SimaPro. SimaPro 8.

(60) OpenLCA. openLCA 1.6.

(61) Wernet, G.; Bauer, C.; Steubing, B.; Reinhard, J.; Moreno-Ruiz, E.; Weidema, B. The Ecoinvent Database Version 3 (Part I): Overview and Methodology. *Int. J. Life Cycle Assess.* **2016**, *21*, 1218–1230.

(62) Pascual-Gonzalez, J.; Guillen-Gosalbez, G.; Mateo-Sanz, J. M.; Jimenez-Esteller, L. Statistical Analysis of the Ecoinvent Database to Uncover Relationships between Life Cycle Impact Assessment Metrics. *J. Clean. Prod.* **2016**, *112*, 359–368.

(63) Bare, J.; Hofstetter, P.; Pennington, D. W.; Udo De Haes, H. A. LCIA Midpoints versus Endpoints: The Sacrifices and Benefits. *Int. J. Life Cycle Assess.* **2000**, *5*, 319–326.

(64) Wolf, M.-A.; Pant, R.; Chomkhamsri, K.; Sala, S.; Pennington, D. Life Cycle Data System (ILCD) Handbook. **2012**.

(65) Hauschild, M. Z.; Goedkoop, M.; Guinée, J.; Heijungs, R.; Huijbregts, M.; Jolliet, O.; Margni, M.; De Schryver, A.; Humbert, S.; Laurent, A.; *et al.* Identifying Best Existing Practice for Characterization Modeling in Life Cycle Impact Assessment. *Int. J. Life Cycle Assess.* **2013**, *18*, 683–697.

(66) Guinée, J. B.; Gorrée M; Heijungs R; Huppes G; Kleijn R; de Koning A; van Oers L; Wegener Sleeswijk A; Suh S; Udo de Haes HA; *et al. Handbook on Life Cycle Assessment: Operational Guide to the ISO Standards*; Kluwer Academic Publishers, 2002.

(67) Goedkoop, M.; Spriensma, R. The Eco-Indicator 99 A Damage Oriented Method for Life Cycle Impact Assessment Methodology Report. **2000**.

(68) Hauschild, M. Z.; Potting, J.; Hertel, O.; Schöpp, W.; Bastrup-Birk, A. Spatial Differentiation in the Characterisation of Photochemical Ozone Formation The EDIP2003 Methodology. *Paris Int J LCA* **2006**, *11*, 72–80.

(69) Steen, B.; Arvidsson, P.; Nobel Gunnar Borg, A.; Louis, S.; Thomas Rydberg, V.; Göran Swan, V.; Enso David Weiner, S. *A Systematic Approach to Environmental Priority Strategies in Product Development (EPS). Version 2000 – General System Characteristics The EPS System The EPS Default Method*; 1999.

(70) Margni, M.; Charles, R.; Humbert, S.; Payet, J.; Rebitzer, G.; Rosenbaum, R.; Jolliet, O. IMPACT 2002+: A New Life Cycle Impact Assessment Methodology. *Int. J. Life Cycle Assess.* **2003**, *8*, 324–330.

(71) Itsubo, N.; Sakagami, M.; Washida, T.; Kokubu, K.; Inaba, A. Weighting across Safeguard Subjects

for LCIA through the Application of Conjoint Analysis. *Int. J. Life Cycle Assess.* **2004**, *9*, 196–205.

(72) Toffoletto, L.; Bulle, C.; Godin, J.; Reid, C.; Deschênes, L. LUCAS - A New LCIA Method Used for a Canadian-Specific Context. *Int. J. Life Cycle Assess.* **2007**, *12*, 93–102.

(73) Huijbregts, M. A. J.; Steinmann, Z. J. N.; Elshout, P. M. F.; Stam, G.; Verones, F.; Vieira, M.; Zijp, M.; Hollander, A.; van Zelm, R. ReCiPe2016: A Harmonised Life Cycle Impact Assessment Method at Midpoint and Endpoint Level. *Int. J. Life Cycle Assess.* **2017**, *22*, 138–147.

(74) Bare, J. TRACI 2.0: The Tool for the Reduction and Assessment of Chemical and Other Environmental Impacts 2.0. *Clean Technol. Environ. Policy* **2011**, *13*, 687–696.

(75) Milà Canals, L.; Romanyà, J.; Cowell, S. J. Method for Assessing Impacts on Life Support Functions (LSF) Related to the Use of "Fertile Land" in Life Cycle Assessment (LCA). *J. Clean. Prod.* **2007**, *15*, 1426–1440.

(76) Carpenter, L. J.; Reimann, S.; Engel, A.; Montzka, S.; Burkholder, J. B.; Clerbaux, C.; Hall, B.; Yvon-Lewis, S. A.; Blake, D. R.; Dorf, M.; *et al.* Update on Ozone-Depleting Substances (ODSs) and Other Gases of Interest to the Montreal Protocol. In *Scientific Assessment of Ozone Depletion: 2014, Global Ozone Research and Monitoring Project-Report No.55*; World Meteorological Organization, Ed.; 2014; p. 416.

(77) Rosenbaum, R. K.; Bachmann, T. M.; Gold, L. S.; Huijbregts, M. A. J.; Jolliet, O.; Juraske, R.; Koehler, A.; Larsen, H. F.; MacLeod, M.; Margni, M.; *et al.* USEtox—the UNEP-SETAC Toxicity Model: Recommended Characterisation Factors for Human Toxicity and Freshwater Ecotoxicity in Life Cycle Impact Assessment. *Int. J. Life Cycle Assess.* **2008**, *13*, 532–546.

(78) Huijbregts, M. A. J.; Rombouts, L. J. A.; Ragas, A. M. J.; van de Meent, D. Human-toxicological Effect and Damage Factors of Carcinogenic and Noncarcinogenic Chemicals for Life Cycle Impact Assessment. *Integr. Environ. Assess. Manag.* **2005**, *1*, 181–244.

(79) Humbert, S. Geographically Differentiated Life-Cycle Impact Assessment of Human Health, University of California, Berkeley, California, USA, 2009.

(80) Rabl Ari; Spadaro Joseph V. The RiskPoll Software, Version 1.052, 2012.

(81) Greco, S. L.; Wilson, A. M.; Spengler, J. D.; Levy, J. I. Spatial Patterns of Mobile Source Particulate Matter Emissions-to-Exposure Relationships across the United States. *Atmos. Environ.* **2007**, *41*, 1011–1025.

(82) van Zelm, R.; Huijbregts, M. A. J.; den Hollander, H. A.; van Jaarsveld, H. A.; Sauter, F. J.; Struijs, J.; van Wijnen, H. J.; van de Meent, D. European Characterization Factors for Human Health Damage of PM10 and Ozone in Life Cycle Impact Assessment. *Atmos. Environ.* **2008**, *42*, 441–453.

(83) Pope, C. A.; Burnett, R. T.; Thun, M. J.; Calle, E. E.; Krewski, D.; Ito, K.; Thurston, G. D.; Thurston, G. D. Lung Cancer, Cardiopulmonary Mortality, and Long-Term Exposure to Fine Particulate Air Pollution. *JAMA* **2002**, *287*, 1132–1141.

(84) Dreicer M; Tort V; Manen P. ExternE, Externalities of Energy, Vol. 5 9 Nuclear, Centr D'étude Sur

l'Evaluation de La Protection Dans Le Domaine 10 Nucléaire (CEPN). In *European Commission DGXII (ed) Science*; 11 Research and development JOULE, 1995.

(85) Frischknecht, R.; Braunschweig, A.; Hofstetter, P.; Suter, P. Human Health Damages due to Ionising Radiation in Life Cycle Impact Assessment. *Environ. Impact Assess. Rev.* **2000**, *20*, 159–189.

(86) Garnier-Laplace, J.; Beaugelin-Seiller, K.; Gilbin, R.; Della-Vedova, C.; Jolliet, O.; Payet, J. A Screening Level Ecological Risk Assessment and Ranking Method for Liquid Radioactive and Chemical Mixtures Released by Nuclear Facilities under Normal Operating Conditions. *Radioprotection* **2009**, *44*, 903–908.

(87) Payet J. Assessing Toxic Impacts on Aquatic Ecosystems in LCA, Ecole Polytechnique Fédérale de Lausanne, 2004.

(88) van Zelm, R.; Huijbregts, M. A. J.; den Hollander, H. A.; van Jaarsveld, H. A.; Sauter, F. J.; Struijs, J.; van Wijnen, H. J.; van de Meent, D. European Characterization Factors for Human Health Damage of PM10 and Ozone in Life Cycle Impact Assessment. *Atmos. Environ.* **2008**, *42*, 441–453.

(89) Seppälä, J.; Posch, M.; Johansson, M.; Hettelingh, J.-P. Country-Dependent Characterisation Factors for Acidification and Terrestrial Eutrophication Based on Accumulated Exceedance as an Impact Category Indicator. *Int. J. Life Cycle Assess.* **2006**, *11*, 403–416.

(90) Posch, M.; Seppälä, J.; Hettelingh, J.-P.; Johansson, M.; Margni, M.; Jolliet, O. The Role of Atmospheric Dispersion Models and Ecosystem Sensitivity in the Determination of Characterisation Factors for Acidifying and Eutrophying Emissions in LCIA. *Int. J. Life Cycle Assess.* **2008**, *13*, 477–486.

(91) Van Zelm, R.; Huijbregts, M. A. J.; van Jaarsveld, H. A.; Reinds, G. J.; de Zwart, D.; Struijs, J.; van de Meent, D. Time Horizon Dependent Characterization Factors for Acidification in Life-Cycle Assessment Based on Forest Plant Species Occurrence in Europe. *Environ. Sci. Technol.* **2007**, *41*, 922–927.

(92) Frischknecht, R.; Steiner, R.; Jungbluth, N. *Methode Der Ökologischen Knappheit – Ökofaktoren 2006*; ö.b.u. und Bundesamt für Umwelt: Bern, 2008.

(93) Taylor, D. The Pharmaceutical Industry and the Future of Drug Development. In *Pharmaceuticals in the Environment*; Hester, R. E.; M., H. R., Eds.; The Royal Society of Chemistry, 2015; pp. 1–33.

(94) Gonzalez, M. A.; Smith, R. L. A Methodology to Evaluate Process Sustainability. *Environ. Prog.* **2003**, *22*, 269–276.

(95) Lapkin, A.; Joyce, L.; Crittenden, B. Framework for Evaluating the "Greenness" of Chemical Processes: Case Studies for a Novel VOC Recovery Technology. *Environemental Sci. Technol.* **2004**, *38*, 5815–5823.

(96) Martins, A. A.; Mata, T. M.; Costa, C. A. V; Sikdar, S. K. Framework for Sustainability Metrics. *Ind. Eng. Chem. Res.* **2007**, *46*, 2962–2973.

(97) Fermeglia, M.; Longo, G.; Toma, L. Computer Aided Design for Sustainable Industrial Processes: Specific Tools and Applications. *AIChE J.* **2009**, *55*, 1065–1078.

(98) Young, D. M.; Cabezas, H. Designing Sustainable Processes with Simulation: The Waste Reduction (WAR) Algorithm. *Comput. Chem. Eng.* **1999**, *23*, 1477–1491.

(99) Othman, M. R.; Repke, J.; Huang, Y.; Wozny, G. A Modular Approach to Sustainability Assessment and Decision Support in Chemical Process Design. *Ind. Eng. Chem. Res.* **2010**, *49*, 7870–7881.

(100) Tugnoli, A.; Santarelli, F.; Cozzani, V. An Approach to Quantitative Sustainability Assessment in the Early Stages of Process Design. *Environ. Sci. Technol.* **2008**, *42*, 4555–4562.

(101) Torres, C. M.; Gadalla, M. A.; Mateo-Sanz, J. M.; Esteller, L. J. Evaluation Tool for the Environmental Design of Chemical Processes. *Ind. Eng. Chem. Res.* **2011**, *50*, 13466–13474.

(102) Shadiya, O. O.; High, K. A. SUSTAINABILITY EVALUATOR: Tool for Evaluating Process Sustainability. *Environ. Prog. Sustain. Energy* **2013**, *32*, 749–761.

(103) Carvalho, A.; Matos, H. A.; Gani, R. SustainPro-A Tool for Systematic Process Analysis, Generation and Evaluation of Sustainable Design Alternatives. *Comput. Chem. Eng.* **2013**, *50*, 8–27.

(104) Babi Deenesh K.; Holtbruegge, J.; Lutze, P.; Gorak, A.; Woodley, J. M.; Gani, R. Sustainable Process Synthesis–intensification. *Comput. Chem. Eng.* **2015**.

(105) Chemmangattuvalappil, N.; Sum, D. N. K.; Elyas, R.; Chen, C.-L.; Chien, I. L.; Lee, H.-Y.; Elms, R. *Chemical Engineering Process Simulation*; Dominic, F., Ed.; Elsevier, 2017.

(106) Schaber, S. D.; Gerogiorgis, D. I.; Ramachandran, R.; Evans, J. M. B.; Barton, P. I.; Trout, B. L. Economic Analysis of Integrated Continuous and Batch Pharmaceutical Manufacturing: A Case Study. *Ind. Eng. Chem. Res.* **2011**, *50*, 10083–10092.

(107) Perry, R. H.; Green, D. W. *Perry's Chemical Engineers' Handbook*; Perry, J. H., Ed.; 8th ed.; McGraw Hill, 2008.

(108) Ulrich, G. D. *A Guide to Chemical Engineering Process Design*; Wiley, 1984.

(109) Vincent, R.; Bonthoux, F.; Mallet, G.; Iparraguire, J.-F.; Rio, S. Méthodologie D'évaluation Simplifiée Du Risque Chimique. *Hygiène sécurité du Trav.* **2005**, *9*, 39–62.

(110) United Nations. *Globally Harmonized System of Classification and Labelling of Chemicals (GHS)*; 2011; Vol. 65.

(111) Smithing, M. P.; Darvas, F. HazardExpert. In; 1992; pp. 191–200.

(112) CompuDrug. Pallas 3.8.

(113) PreADMET https://preadmet.bmdrc.kr/ (accessed Sep 15, 2017).

(114) Sanguinetti, M. C.; Tristani-Firouzi, M. hERG Potassium Channels and Cardiac Arrhythmia. *Nature* **2006**, *440*, 463–469.

(115) Description of PreADMET https://preadmet.bmdrc.kr/description-of-preadmet/ (accessed Sep 15, 2017).

(116) In, Y.; Lee, S. K.; Kim, P. J.; No, K. T. Prediction of Acute Toxicity to Fathead Minnow by Local Model Based QSAR and Global QSAR Approaches. *Bull. Korean Chem. Soc.* **2012**, *33*, 613–619.

(117) Ames, B. N.; Durston, W. E.; Yamasaki, E.; Lee, F. D. Carcinogens Are Mutagens: A Simple Test

System Combining Liver Homogenates for Activation and Bacteria for Detection. *Proc. Natl. Acad. Sci. U. S. A.* **1973**, *70*, 2281–2285.

(118) Klamt, A.; Eckert, F.; Diedenhofen, M. Prediction of Soil Sorption Coefficients with a Conductor-like Screening Model for Real Solvents. *Environ. Toxicol. Chem.* **2002**, *21*, 2562–2566.

(119) Mayo-Bean, K.; Moran-Bruce, K.; Meylan, W.; Ranslow, P.; Lock, M.; Nabholz, J. V.; Runnen, J. Von; Cassidy, L. M.; Tunkel, J. ECOlogical Structure-Activity Relationship Model (ECOSAR) Class Program ESTIMATING TOXICITY OF INDUSTRIAL CHEMICALS TO AQUATIC ORGANISMS USING THE ECOSAR (ECOLOGICAL STRUCTURE-ACTIVITY RELATIONSHIP) CLASS PROGRAM Version 2.0. **2017**.

(120) US EPA, O. EPI Suite™-Estimation Program Interface.

(121) Grisoni, F.; Consonni, V.; Villa, S.; Vighi, M.; Todeschini, R. QSAR Models for Bioconcentration: Is the Increase in the Complexity Justified by More Accurate Predictions? *Chemosphere* **2015**, *127*, 171–179.

(122) ECETOC. *The Role of Bioaccumulation in Environmental Risk Assessment: The Aquatic Environment and Related Food Webs*; 1995.

(123) European Commission (EC). *REGULATION (EC) No 1907/2006 OF THE EUROPEAN PARLIAMENT AND OF THE COUNCIL of 18 December 2006 Concerning the Registration, Evaluation, Authorisation and Restriction of Chemicals (REACH), Establishing a European Chemicals Agency, Amending Directive 1999/4*; 2007; pp. 1–849.

(124) Gissi, A.; Gadaleta, D.; Floris, M.; Olla, S.; Carotti, A.; Novellino, E.; Benfenati, E.; Nicolotti, O. An Alternative QSAR-Based Approach for Predicting the Bioconcentration Factor for Regulatory Purposes. *ALTEX* **2014**, *31*, 23–36.

(125) Klamt, A. Estimation of Gas-Phase Hydroxyl Radical Rate Constants of Oxygenated Compounds Based on Molecular Orbital Calculations. *Chemosphere* **1996**, *32*, 717–726.

(126) Charnes, A.; Cooper, W. W.; Rhodes, E. Measuring the Efficiency of Decision Making Units. *Eur. J. Oper. Res.* **1978**, *2*, 429–444.

(127) Ren, J.; Tan, S.; Dong, L.; Mazzi, A.; Scipioni, A.; Sovacool, B. K. Determining the Life Cycle Energy Efficiency of Six Biofuel Systems in China: A Data Envelopment Analysis. *Bioresour. Technol.* **2014**, *162*, 1–7.

(128) Vázquez-Rowe, I.; Iribarren, D. Review of Life-Cycle Approaches Coupled with Data Envelopment Analysis: Launching the CFP + DEA Method for Energy Policy Making. *Sci. World J.* **2015**.

(129) Lorenzo-Toja, Y.; Vazquez-Rowe, I.; Chenel, S.; Marin-Navarro, D.; Moreira, M. T.; Feijoo, G. Eco-Efficiency Analysis of Spanish WWTPs Using the LCA+DEA Method. *Water Res.* **2015**, *68*, 637–650.

(130) Zhou, P.; Ang, B. W.; Poh, K. L. A Survey of Data Envelopment Analysis in Energy and Environmental Studies. *Eur. J. Oper. Res.* **2008**, *189*, 1–18.

(131) Zhu, J. *Data Envelopment Analysis: A Handbook of Empirical Studies and Applications*; Springer

Science + Business Media: New York, 2016.

(132) Limleamthong, P.; Gonzalez-Miquel, M.; Papadokonstantakis, S.; Papadopoulos, A. I.; Seferlis, P.; Guillén-Gosálbez, G. Multi-Criteria Screening of Chemicals Considering Thermodynamic and Life Cycle Assessment Metrics via Data Envelopment Analysis: Application to CO2 Capture. *Green Chem.* **2016**, *18*, 6468–6481.

(133) Cooper, W. W.; Seiford, L. M.; Tone, K. *Data Envelopment Analysis A Comprehensive Text with Models, Applications, References and DEA-Solver Software*; Springer Science & Business Media, 2006.

(134) Banker, R. D. Estimating Most Productive Scale Size Using Data Envelopment Analysis. *Eur. J. Oper. Res.* **1984**, *17*, 35–44.

(135) Cook, W. D.; Seiford, L. M. Data Envelopment Analysis (DEA) – Thirty Years on. *Eur. J. Oper. Res.* **2009**, *192*, 1–17.

(136) Leo, A.; Hansch, C.; Elkins, D. Partition Coefficients and Their Uses. *Chem. Rev.* **1971**, *71*, 525–616.

(137) Shargel, L.; Yu, A. B. C.; Wu-Pong, S. Chapter 10: Physiological Drug Distribution and Protein Binding. In *Applied Biopharmaceutics & Pharmacokinetics*; McGraw-Hill Medical: New York, 2012; p. 811.

(138) Geyer, H.; Politzki, G.; Freitag, D. Prediction of Ecotoxicological Behaviour of Chemicals: Relationship between N-Octanol/water Partition Coefficient and Bioaccumulation of Organic Chemicals by Alga Chlorella. *Chemosphere* **1984**, *13*, 269–284.

(139) Hermens, J. L. M.; de Bruijn, J. H. M.; Brooke, D. N. The Octanol-Water Partition Coefficient: Strengths and Limitations. *Environ. Toxicol. Chem.* **2013**, *32*, 732–733.

(140) Ulas, S.; Diwekar, U. M. Efficient Molecular Simulations for Environmentally Benign Processes. *Mol. Simul.* **2006**, *32*, 315–329.

(141) Tewarl, Y. B.; Mlller, M. M.; Waslk, S. P.; Martlret, D. E. Aqueous Solubility and Octano-Water Partition Coefficient of Organic Compounds at 25.0 C. *J. Chem. Eng. Data* **1962**, *27*, 451–454.

(142) Tetko, I. V; Poda, G. I. Application of ALOGPS 2.1 to Predict Log D Distribution Coefficient for Pfizer Proprietary Compounds. *J. Med. Chem.* **2004**, *47*, 5601–5604.

(143) Sijm, D.; Schuurmann, G.; De Vries, P.; Opperhuizen, A. Aqueous Solubility, Octanol Solubility, and Octanol Water Partition Coefficient of Nine Hydrophobic Dyes ENV TOX CH, 18(6), 1999, Pp. 1109-1117. *Environ. Toxicol. Chem.* **1999**, *18*, 1109–1117.

(144) Lang, B. E. Solubility of Water in Octan-1-Ol from (275 to 369) K. *J. Chem. Eng. Data* **2012**, *57*, 2221–2226.

(145) Ren, S.; Frymier, D. P.; Schultz, T. W. An Exploratory Study of the Use of Multivariate Techniques to Determine Mechanisms of Toxic Action. *Ecotoxicol. Environ. Saf.* **2003**, *55*, 86–97.

(146) Lin, Z.; Shi, P.; Gao, S.; Wang, L.; Yu, H. Use of Partition Coefficients to Predict Mixture Toxicity. *Water Res.* **2003**, *37*, 2223–2227.

(147) Henchoz, Y.; Guillarme, D.; Rudaz, S.; Veuthey, J.-L.; Carrupt, P.-A. High-Throughput Log *P* Determination by Ultraperformance Liquid Chromatography: A Convenient Tool for Medicinal Chemists. *J. Med. Chem.* **2008**, *51*, 396–399.

(148) Qiao, Y.; Xia, S.; Ma, P. Octanol/Water Partition Coefficient of Substituted Benzene Derivatives Containing Halogens and Carboxyls: Determination Using the Shake-Flask Method and Estimation Using the Fragment Method. *J. Chem. Eng. Data* **2008**, *53*, 280–282.

(149) Umland, J. B. The Determination of 1-Octanol/water Partition Ratios: An Organic Chemistry Laboratory Experiment. *J. Chem. Educ.* **1983**, *60*, 1081.

(150) Paschke, A.; Neitzel, P. L.; Walther, W.; Schüürmann, G. Octanol/Water Partition Coefficient of Selected Herbicides: Determination Using Shake-Flask Method and Reversed-Phase High-Performance Liquid Chromatography. *J. Chem. Eng. Data* **2004**, *49*, 1639–1642.

(151) Brooke, D. N.; Dobbs, A. J.; Williams, N. Octanol:water Partition Coefficients (P): Measurement, Estimation, and Interpretation, Particularly for Chemicals with P Greater than 10^5. *Ecotoxicol. Environ. Saf.* **1986**, *11*, 251–260.

(152) Organisation for Economic Cooperation and Development (OECD). *Partition Coefficient (1-Octanol/water): Slow-Stirring Method. Test Guideline 123*; Paris, France, 2006.

(153) Tolls, J.; Bodo, K.; De Felip, E.; Dujardin, R.; Kim, Y. H.; Moeller-Jensen, L.; Mullee, D.; Nakajima, A.; Paschke, A.; Pawliczek, J.-B.; *et al.* Slow-Stirring Method for Determining the N-Octanol/water Partition Coefficient (Pow) for Highly Hydrophobic Chemicals: Performance Evaluation in a Ring Test. *Environ. Toxicol. Chem.* **2003**, *22*, 1051–1057.

(154) Scifinder Platform https://www.cas.org/ (accessed Jul 1, 2017).

(155) Åberg, K. M.; Lyubartsev, A. P.; Jacobsson, S. P.; Laaksonen, A.; Magnus, K.; Berg, Å. Determination of Solvation Free Energies by Adaptive Expanded Ensemble Molecular Dynamics. *J. Chem. Phys. J. Chem. Phys. J. Chem. Phys. J. Chem. Phys. J. Chem. Phys. J. Chem. Phys.* **2004**, *1201*, 1776–154111.

(156) Wikipedia https://en.wikipedia.org/wiki/Thermodynamic_integration (accessed Sep 25, 2017).

(157) Best, S. A.; Merz, K. M.; Reynolds, C. H. Free Energy Perturbation Study of Octanol/Water Partition Coefficients: Comparison with Continuum GB/SA Calculations. *J. Phys. Chem. B* **1999**, *103*, 714–726.

(158) Zwanzig, R. W. High-Temperature Equation of State by a Perturbation Method. I. Nonpolar Gases. *J. Chem. Phys.* **1954**, *22*, 1420–1426.

(159) DeBolt, S. E.; Kollman, P. A. Investigation of Structure, Dynamics, and Solvation in 1-Octanol and Its Water-Saturated Solution: Molecular Dynamics and Free-Energy Perturbation Studies. *J. Am. Chem. Soc.* **1995**, *117*, 5316–5340.

(160) Torrie, G. M.; Valleau, J. P. Nonphysical Sampling Distributions in Monte Carlo Free-Energy Estimation: Umbrella Sampling. *J. Comput. Phys.* **1977**, *23*, 187–199.

(161) Mezei, M. Adaptive Umbrella Sampling: Self-Consistent Determination of the Non-Boltzmann Bias. *J. Comput. Phys.* **1987**, *68*, 237–248.

(162) Lyubartsev, A. P.; Martsinovski, A. A.; Shevkunov, S. V.; Vorontsov-Velyaminov, P. N. New Approach to Monte Carlo Calculation of the Free Energy: Method of Expanded Ensembles. *J. Chem. Phys.* **1992**, *96*, 1776–1783.

(163) Chen, B.; Siepmann, J. I. Partitioning of Alkane and Alcohol Solutes between Water and (Dry or Wet) 1-Octanol. *J. Am. Chem. Soc.* **2000**, *122*, 6464–6467.

(164) Martin, M. G.; Siepmann, J. I. Calculating Gibbs Free Energies of Transfer from Gibbs Ensemble Monte Carlo Simulations. *Theor. Chem. Accounts Theory, Comput. Model. (Theoretica Chim. Acta)* **1998**, *99*, 347–350.

(165) Nantasenamat, C.; Isarankura-Na-Ayudhya, C.; Naenna, T.; Prachayasittikul, V. A PRACTICAL OVERVIEW OF QUANTITATIVE STRUCTURE-ACTIVITY RELATIONSHIP. *EXCLI J.* **2009**, *8*, 74–88.

(166) *Handbook of Molecular Descriptors*; Todeschini, R.; Consonni, V., Eds.; Methods and Principles in Medicinal Chemistry; Wiley-VCH Verlag GmbH: Weinheim, Germany, 2000.

(167) Karelson, M.; Lobanov, V. S.; Katritzky, A. R. Quantum-Chemical Descriptors in QSAR/QSPR Studies. *Chem. Rev.* **1996**, *96*, 1027–1043.

(168) Tehrany, E. A.; Fournier, F.; Desobry, S. Simple Method to Calculate Octanol–water Partition Coefficient of Organic Compounds. *J. Food Eng.* **2004**, *64*, 315–320.

(169) Bodor, N.; Gabanyi, Z.; Wong, C.-K. A New Method for the Estimation of Partition Coefficient. *J. Am. Chem. SOC* **1989**, *1*, 3783–3786.

(170) Eisfeld, W.; Maurer, G. Study on the Correlation and Prediction of Octanol/Water Partition Coefficients by Quantum Chemical Calculations. *J. Phys. Chem. B* **1999**, *103*, 5716–5729.

(171) Amat, L.; Carbo-Dorca, R.; Ponec, R. Molecular Quantum Similarity Measures as an Alternative to Log P Values in QSAR Studies. *J. Comput. Chem.* **1998**, *19*, 1575–1583.

(172) Xing, L.; Glen, R. C. Novel Methods for the Prediction of logP, pK a , and logD. *J. Chem. Inf. Comput. Sci.* **2002**, *42*, 796–805.

(173) Klopman, G.; Iroff, L. D. Calculation of Partition Coefficients by the Charge Density Method. *J. Comput. Chem.* **1981**, *2*, 157–160.

(174) Yao, X.; Wang, Y.; Zhang, X.; Zhang, R.; Liu, M.; Hu, Z.; Fan, B. Radial Basis Function Neural Network-Based QSPR for the Prediction of Critical Temperature. *Chemom. Intell. Lab. Syst.* **2002**, *62*, 217–225.

(175) Galvão, R. K. H.; Araújo, M. C. U.; Fragoso, W. D.; Silva, E. C.; José, G. E.; Soares, S. F. C.; Paiva, H. M. A Variable Elimination Method to Improve the Parsimony of MLR Models Using the Successive Projections Algorithm. *Chemom. Intell. Lab. Syst.* **2008**, *92*, 83–91.

(176) Gramatica, P.; Chirico, N.; Papa, E.; Cassani, S.; Kovarich, S. QSARINS: A New Software for the Development, Analysis, and Validation of QSAR MLR Models. *J. Comput. Chem.* **2013**, *34*, 2121–

2132.

(177) Suzuki, T.; Ohtaguchi, K.; Koide, K. Correlations between Octanol/water Partition Coefficients and Intrinsic Molecular Properties of Simple Organic Compounds. *J. Chem. Eng. JAPAN* **1993**, *26*, 114–115.

(178) Duprat, A. F.; Huynh, T.; Dreyfus, G. Toward a Principled Methodology for Neural Network Design and Performance Evaluation in QSAR. Application to the Prediction of LogP. *J. Chem. Inf. Comput. Sci.* **1998**, *38*, 586–594.

(179) Jarmo J. Huuskonen; David J. Livingstone; Igor V. Tetko. Neural Network Modeling for Estimation of Partition Coefficient Based on Atom-Type Electrotopological State Indices. *J. Chem. Inf. Comput. Sci.* **2000**, *40*, 947–955.

(180) Tetko, I. V.; Tanchuk, V. Y.; Villa, A. E. P. Prediction of N-Octanol/Water Partition Coefficients from PHYSPROP Database Using Artificial Neural Networks and E-State Indices. *J. Chem. Inf. Comput. Sci.* **2001**, *41*, 1407–1421.

(181) Cense, J. M.; B., D.; J.J., L.; G., R. Neural Networks Prediction of Partition Coefficients. *Chemom. Intell. Lab. Syst.* **1994**, *23*, 301–308.

(182) Wold, S.; Sjöström, M.; Eriksson, L. PLS-Regression: A Basic Tool of Chemometrics. *Chemom. Intell. Lab. Syst.* **2001**, *58*, 109–130.

(183) Öberg, T. A QSAR for Baseline Toxicity: Validation, Domain of Application, and Prediction. *Chem. Res. Toxicol.* **2004**, *17*, 1630–1637.

(184) Axelsson, C.; Skidmore, A. K.; Schlerf, M.; Fauzi, A.; Verhoef, W. Hyperspectral Analysis of Mangrove Foliar Chemistry Using PLSR and Support Vector Regression. *Int. J. Remote Sens.* **2013**, *34*, 1724–1743.

(185) Wold, S.; Sjöström, M.; Eriksson, L. PLS-Regression: A Basic Tool of Chemometrics. *Chemom. Intell. Lab. Syst.* **2001**, *58*, 109–130.

(186) Li, H.; Liang, Y.; Xu, Q. Support Vector Machines and Its Applications in Chemistry. *Chemom. Intell. Lab. Syst.* **2009**, *95*, 188–198.

(187) ZHAO, C.; ZHANG, H.; ZHANG, X.; LIU, M.; HU, Z.; FAN, B. Application of Support Vector Machine (SVM) for Prediction Toxic Activity of Different Data Sets. *Toxicology* **2006**, *217*, 105–119.

(188) Yang, S.-S.; Lu, W.-C.; Gu, T.-H.; Yan, L.-M.; Li, G.-Z. QSPR Study of N -Octanol/Water Partition Coefficient of Some Aromatic Compounds Using Support Vector Regression. *QSAR Comb. Sci.* **2009**, *28*, 175–182.

(189) Smola, A. J.; Sc Olkopf, B. A Tutorial on Support Vector Regression. *Stat. Comput.* **2004**, *14*, 199–222.

(190) Ferreira, C. Gene Expression Programming in Problem Solving. In *Soft Computing and Industry – Recent Applications*; Springer-Verlag, 2002; pp. 635–654.

(191) Liu, P.; Long, W. Current Mathematical Methods Used in QSAR/QSPR Studies. *Int. J. Mol. Sci.* **2009**,

10, 1978–1998.

(192) Han, X. R.; Li, X. C.; Si, H. Z.; Ge, C. Z.; Gao, H.; Zhai, H. L. QSAR Study of the Anti-Cancer Activity of 38 Compounds in Different Cancer Cell Lines Based on Gene Expression Programming. *Adv. Mater. Res.* **2013**, *850–851*, 1291–1294.

(193) Song, F.; Zhang, A.; Liang, H.; Cui, L.; Li, W.; Si, H.; Duan, Y.; Zhai, H. QSAR Study for Carcinogenic Potency of Aromatic Amines Based on GEP and MLPs. *Int. J. Environ. Res. Public Health* **2016**, *13*, 1141.

(194) Si, H. Z.; Zhang, K. J.; Hu, Z. D.; Fan, B. T. QSAR Model for Prediction Capacity Factor of Molecular Imprinting Polymer Based on Gene Expression Programming. *QSAR Comb. Sci.* **2007**, *26*, 41–50.

(195) Friedman, J. H.; Stuetzle, W. Projection Pursuit Regression. *J. Am. Stat. Assoc.* **1981**, *76*, 817–823.

(196) Yuan, J.; Xie, C.; Zhang, T.; Sun, J.; Yuan, X.; Yu, S.; Zhang, Y.; Cao, Y.; Yu, X.; Yang, X.; *et al.* Linear and Nonlinear Models for Predicting Fish Bioconcentration Factors for Pesticides. *Chemosphere* **2016**, *156*, 334–340.

(197) Doucet, J. P.; Papa, E.; Doucet-Panaye, A.; Devillers, J. QSAR Models for Predicting the Toxicity of Piperidine Derivatives against *Aedes Aegypti*. *SAR QSAR Environ. Res.* **2017**, *28*, 451–470.

(198) Ren, Y.; Liu, H.; Yao, X.; Liu, M. Prediction of Ozone Tropospheric Degradation Rate Constants by Projection Pursuit Regression. *Anal. Chim. Acta* **2007**, *589*, 150–158.

(199) Lei, B.; Ma, Y.; Li, J.; Liu, H.; Yao, X.; Gramatica, P. Prediction of the Adsorption Capability onto Activated Carbon of a Large Data Set of Chemicals by Local Lazy Regression Method. *Atmos. Environ.* **2010**, *44*, 2954–2960.

(200) Lu, J.; Peng, J.; Wang, J.; Shen, Q.; Bi, Y.; Gong, L.; Zheng, M.; Luo, X.; Zhu, W.; Jiang, H.; *et al.* Estimation of Acute Oral Toxicity in Rat Using Local Lazy Learning. *J. Cheminform.* **2014**, *6*, 26.

(201) Lu, J.; Lu, D.; Zhang, X.; Bi, Y.; Cheng, K.; Zheng, M.; Luo, X. Estimation of Elimination Half-Lives of Organic Chemicals in Humans Using Gradient Boosting Machine. *Biochim. Biophys. Acta - Gen. Subj.* **2016**, *1860*, 2664–2671.

(202) Klamt, A. The COSMO and COSMO-RS Solvation Models. *Wiley Interdiscip. Rev. Comput. Mol. Sci.* **2011**, *1*, 699–709.

(203) Tomasi, J.; Mennucci, B.; Cammi, R. Quantum Mechanical Continuum Solvation Models. *Chem. Rev.* **2005**, *105*, 2999–3093.

(204) Klamt, A. Conductor-like Screening Model for Real Solvents: A New Approach to the Quantitative Calculation of Solvation Phenomena. *J. Phys. Chem* **1995**, *99*, 2224–2235.

(205) Geerlings, P.; De Proft, F.; Langenaeker, W. Conceptual Density Functional Theory. *Chem. Rev.* **2003**, *103*, 1793–1873.

(206) Machatha, S. G.; Yalkowsky, S. H. Comparison of the Octanol/water Partition Coefficients Calculated by ClogP??, ACDlogP and KowWin?? To Experimentally Determined Values. *Int. J. Pharm.* **2005**, *294*, 185–192.

(207) Grisoni, F.; Consonni, V.; Villa, S.; Vighi, M.; Todeschini, R. QSAR Models for Bioconcentration: Is the Increase in the Complexity Justified by More Accurate Predictions? *Chemosphere* **2015**, *127*, 171–179.

(208) DrugBank https://www.drugbank.ca/ (accessed Sep 8, 2017).

(209) ChemSpider | Search and share chemistry http://www.chemspider.com/ (accessed Sep 9, 2017).

(210) The PubChem Project https://pubchem.ncbi.nlm.nih.gov/ (accessed Sep 8, 2017).

(211) Bolton, E. E.; Chen, J.; Kim, S.; Han, L.; He, S.; Shi, W.; Simonyan, V.; Sun, Y.; Thiessen, P. A.; Wang, J.; *et al.* PubChem3D: A New Resource for Scientists. *J. Cheminform.* **2011**, *3*, 32.

(212) Crawley, M. J. *Statistics, An Introduction Using R*; Wiley, Ed.; 2nd ed.; 2015.

(213) EPA iCSS ToxCast Dashboard https://actor.epa.gov/dashboard/ (accessed Sep 15, 2017).

(214) Al-Majed, A.; Bakheit, A. H. H.; Abdel Aziz, H. A.; Alharbi, H.; Al-Jenoobi, F. I. Pioglitazone. In *Profiles of drug substances, excipients, and related methodology*; 2016; Vol. 41, pp. 379–438.

(215) Janani, C.; Ranjitha Kumari, B. PPAR Gamma Gene – A Review. *Diabetes Metab. Syndr. Clin. Res. Rev.* **2015**, *9*, 46–50.

(216) U.S. Food and Drug Administration. *Approved Drug Products With Therapeutic Equivalence Evaluations, 37th Edition, "Orange Book";* 2017.

(217) Hsiao, F.-Y.; Hsieh, P.-H.; Huang, W.-F.; Tsai, Y.-W.; Gau, C.-S. Risk of Bladder Cancer in Diabetic Patients Treated with Rosiglitazone or Pioglitazone: A Nested Case-Control Study. *Drug Saf.* **2013**, *36*, 643–649.

(218) Neumann, A.; Weill, A.; Ricordeau, P.; Fagot, J. P.; Alla, F.; Allemand, H. Pioglitazone and Risk of Bladder Cancer among Diabetic Patients in France: A Population-Based Cohort Study. *Diabetologia* **2012**, *55*, 1953–1962.

(219) Tuccori, M.; Filion, K. B.; Yin, H.; Yu, O. H.; Platt, R. W.; Azoulay, L. Pioglitazone Use and Risk of Bladder Cancer: Population Based Cohort Study. *BMJ* **2016**, *352*.

(220) Korhonen, P.; Heintjes, E. M.; Williams, R.; Hoti, F.; Christopher, S.; Majak, M.; Kool-Houweling, L.; Strongman, H.; Linder, M.; Dolin, P.; *et al.* Pioglitazone Use and Risk of Bladder Cancer in Patients with Type 2 Diabetes: Retrospective Cohort Study Using Datasets from Four European Countries. *BMJ* **2016**, *354*.

(221) Strongman, H.; Korhonen, P.; Williams, R.; Bahmanyar, S.; Hoti, F.; Christopher, S.; Majak, M.; Kool-Houweling, L.; Linder, M.; Dolin, P.; *et al.* Pioglitazone and Risk of Mortality in Patients with Type 2 Diabetes: Results from a European Multidatabase Cohort Study. *BMJ Open Diab Res Care* **2017**, *5*.

(222) Lewis, J. D.; Habel, L. A.; Quesenberry, C. P.; Strom, B. L.; Peng, T.; Hedderson, M. M.; Ehrlich, S. F.; Mamtani, R.; Bilker, W.; Vaughn, D. J.; *et al.* Pioglitazone Use and Risk of Bladder Cancer and Other Common Cancers in Persons With Diabetes. *JAMA* **2015**, *314*, 265.

(223) Harrington, P. J. *Pharmaceutical Process Chemistry for Synthesis: Rethinking the Routes to Scale-Up*;

Wiley, 2011.

(224) Pioglitazone HCl - Uses, DMF, Dossier, Manufacturer, Supplier, Licensing, Distributer, Prices, News, GMP https://www.pharmacompass.com/active-pharmaceutical-ingredients/pioglitazone-hcl (accessed Jul 20, 2017).

(225) Meguro, K.; Fujita, T.; Hatanaka, C.; Ooi, S. Method for Producing Thiazolidinedione Derivatives. US 4,812,570, 1989.

(226) Sohda, T.; Momose, Y.; Meguro, K.; Kawamatsu, Y.; Sugiyama, Y.; Ikeda, H. Studies on Antidiabetic Agents. Synthesis and Hypoglycemic Activity of 5-[4-(Pyridylalkoxy)benzyl]-2,4-Thiazolidinediones. *Arzneimittel-Forschung/Drug Res.* **1990**, *40*.

(227) Saito, Y.; Mizufune, H.; Yamashita, M. Production Of Benzaldehyde Compounds. US 6,100,403, 2000.

(228) Rajendra, G. L.; Parag, N. G.; Maya, J. S.; Janak, R. S. Improved Process for Preparation of Thiazolidinedione Derivatives. WO2004024059 A2, 2004.

(229) Momose, Y.; Meguro, K.; Ikeda, H.; Hatanaka, C.; Oi, S.; Sohda, T. Studies on Antidiabetic Agents. X. Synthesis and Biological Activities of Pioglitazone and Related Compounds. *Chem. Pharm. Bull.* **1991**, *39*, 1440–1447.

(230) Suri, S.; Sarin, G. S. An Improved Process for the Production of Derivatives of Thiazolidinediones and Their Precursors. WO2006035459 A1, 2006.

(231) Madivada, L. R.; Anumala, R. R.; Gilla, G.; Alla, S.; Charagondla, K.; Kagga, M.; Bhattacharya, A.; Bandichhor, R. An Improved Process for Pioglitazone and Its Pharmaceutically Acceptable Salt. *Org. Process Res. Dev.* **2009**, *13*, 1190–1194.

(232) Huber, J. E. Reduction Method for Substituted 5-Methylene-Thiazolidinediones. U.S. Patent 5,585,495, 1996.

(233) Mohanty, S.; Talasila, S.; Roy, A. K.; Karmakar, A. C. Development of an Improved and Scalable Process for 2-(5-Ethylpyridin-2-Yl) Ethan-1-Ol: Solvent-Free Reaction and Recycling of the Starting Material 5-Ethyl-2-Picoline. *Org. Process Res. Dev.* **2014**, *18*, 168–173.

(234) Pandey, B. NOVEL PROCESS TO PREPARE PIOGLITAZONE VIA SEVERAL NOVEL INTERMEDIATES. US 20140088127 A1, 2014.

(235) Oba, Y.; Tanaka, K.; Oshiki, H.; Kamizono, H. Preparation of High-Purity 5-[4-[2-(5-Ethyl-2-Pyridyl)ethoxy]benzyl]-2,4-Thiazolidinedione Hydrochloride and Its Intermediate. JP,2010–105943,A, 2010.

(236) Boussofiane, A.; Dyson, R. G.; Thanassoulis, E. Applied Data Envelopment Analysis. *Eur. J. Oper. Res.* **1991**, *52*, 1–15.

(237) GAMS Development Corporation. *General Algebraic Modeling System (GAMS) Release 24.4.5*; Washington, DC, USA, 2015.

(238) GAMS Development Corporation. *GAMS - The Solver Manuals, GAMS Release 24.4.5*; Washington,

DC, USA, 2015.

(239) Anastopoulos, G.; Zannikou, Y.; Stournas, S.; Kalligeros, S. Transesterification of Vegetable Oils with Ethanol and Characterization of the Key Fuel Properties of Ethyl Esters. *Energies* **2009**, *2*, 362–376.

(240) Zhang, Y.; Dube, M. A.; McLean, D. D.; Kates, M. Biodiesel Production from Waste Cooking Oil: 1. Process Design and Technological Assessment. *Bioresour. Technol.* **2003**, *89*, 1–16.

(241) Che Man, Y. B.; Haryati, T.; Ghazali, H. M.; Asbi, B. A. Composition and Thermal Profile of Crude Palm Oil and Its Products. *J. Am. Oil Chem. Soc.* **1999**, *76*, 237–242.

(242) Neste https://www.neste.com/en/corporate-info/investors/market-data/biodiesel-prices-sme-fame (accessed Aug 18, 2017).

(243) Indexmundi http://www.indexmundi.com/commodities/?commodity=palm-oil (accessed Aug 20, 2017).

(244) Weller, H. Colloidal Semiconductor Q-Particles: Chemistry in the Transition Region Between Solid State and Molecules. *Angew. Chemie Int. Ed. English* **1993**, *32*, 41–53.

(245) Medintz, I. L.; Uyeda, H. T.; Goldman, E. R.; Mattoussi, H. Quantum Dot Bioconjugates for Imaging, Labelling and Sensing. *Nat. Mater.* **2005**, *4*, 435–446.

(246) Kim, G.-H.; García de Arquer, F. P.; Yoon, Y. J.; Lan, X.; Liu, M.; Voznyy, O.; Yang, Z.; Fan, F.; Ip, A. H.; Kanjanaboos, P.; *et al.* High-Efficiency Colloidal Quantum Dot Photovoltaics via Robust Self-Assembled Monolayers. *Nano Lett.* **2015**, *15*, 7691–7696.

(247) Zhao, J.; Holmes, M. A.; Osterloh, F. E. Quantum Confinement Controls Photocatalysis: A Free Energy Analysis for Photocatalytic Proton Reduction at CdSe Nanocrystals. *ACS Nano* **2013**, *7*, 4316–4325.

(248) Coe, S.; Woo, W.-K.; Bawendi, M.; Bulović, V. Electroluminescence from Single Monolayers of Nanocrystals in Molecular Organic Devices. *Nature* **2002**, *420*, 800–803.

(249) Kale, P. G.; Solanki, C. S. Silicon Quantum Dot Solar Cell Using Top-down Approach. *Int. Nano Lett.* **2015**, *5*, 61–65.

(250) Murray, C. B.; Norris, D. J.; Bawendi, M. G. Synthesis and Characterization of Nearly Monodisperse CdE (E = Sulfur, Selenium, Tellurium) Semiconductor Nanocrystallites. *J. Am. Chem. Soc.* **1993**, *115*, 8706–8715.

(251) Slejko, E. A. Functional Heterostructures Based on Colloidal Nanocrystals, Università di Trieste, 2017.