Universidade Federal do Pará

Instituto de Ciências Exatas e Naturais

Faculdade de Computação

Igor Matheus Souza Moreira

On Reducing the Dimensionality of Small Molecule Data for Visual-Exploratory Analysis in Human Intestinal Absorption Prediction

Belém, PA, Brazil

February 2022

Igor Matheus Souza Moreira

On Reducing the Dimensionality of Small Molecule Data for Visual-Exploratory Analysis in Human Intestinal Absorption Prediction

Course completion work presented to the Faculdade de Computação of the Instituto de Ciências Exatas e Naturais as one of the requirements to complete the computer science bachelor's degree program offered by Universidade Federal do Pará.

Universidade Federal do Pará

Instituto de Ciências Exatas e Naturais

Faculdade de Computação

Supervisor: Prof. Dr. Claudomiro de Souza de Sales Junior

Co-supervisor: Ing. Msc. Ewerton Cristhian Lima de Oliveira

Belém, PA, Brazil

February 2022

Igor Matheus Souza Moreira

On Reducing the Dimensionality of Small Molecule Data for Visual-Exploratory Analysis in Human Intestinal Absorption Prediction/ Igor Matheus Souza Moreira. – Belém, PA, Brazil, February 2022-

137p. : il. (some color.) ; 30 cm.

Supervisor: Prof. Dr. Claudomiro de Souza de Sales Junior

Course Completion Work – Universidade Federal do Pará
Instituto de Ciências Exatas e Naturais
Faculdade de Computação, February 2022.

1. Chemoinformatics. 2. Computational pharmaceutics. 3. Dimensionality reduction. 4. Drug discovery and development. 5. Feature extraction. 6. Human intestinal absorption. 7. Machine learning. I. Prof. Dr. Claudomiro de Souza de Sales Junior. II. *Universidade Federal do Pará*. III. *Faculdade de Computação*. IV. On Reducing the Dimensionality of Small Molecule Data for Visual-Exploratory Analysis in Human Intestinal Absorption Prediction

UNIVERSIDADE FEDERAL DO PARÁ INSTITUTO DE CIÊNCIAS EXATAS E NATURAIS FACULDADE DE COMPUTAÇÃO

ATA DA DEFESA PÚBLICA DO TRABALHO DE CONCLUSÃO DO CURSO DE CIÊNCIA DA COMPUTAÇÃO REALIZADO PELO DISCENTE IGOR MATHEUS SOUZA MOREIRA.

Aos 21 de fevereiro de 2022, às dez horas e trinta minutos, realizou-se, de forma remota, por meio da plataforma Google Meet, a sessão de Defesa Pública do Trabalho de Conclusão de Curso intitulado 'ON REDUCING THE DIMENSIONALITY OF SMALL MOLECULE DATA FOR VISUAL-EXPLORATORY ANALYSIS IN HUMAN INTESTINAL ABSORPTION PREDICTION', apresentado pelo discente IGOR MATHEUS SOUZA MOREIRA. A sessão foi instalada às às dez horas e trinta minutos pelo Prof. Dr. Claudomiro de Souza de Sales Junior. A referida banca foi constituída pelos seguintes membros: Prof. Dr. Claudomiro de Souza de Sales Junior (ORIENTADOR), Ing. Msc. Ewerton Cristhian Lima de Oliveira (COORIENTADOR), Profª. Drª. Regiane Silva Kawasaki Francês (AVALIADORA), Prof. Dr. Reginaldo Cordeiro dos Santos Filho (AVALIADOR), Dr. Caio Marcos Flexa Rodrigues (AVALIADOR), Prof. Dr. Kauê Santana da Costa (AVALIADOR). A Banca Examinadora, após a exposição do mencionado Trabalho pelo discente, passou a arguí-lo. E nada mais havendo a tratar, o presidente deu por encerrada a Defesa do Trabalho, agradecendo a presença de todos, e para constar a legitimidade do que foi deliberado, lavrou-se a presente ata que após lida, será assinada pelos membros presentes na reunião. Belém, vinte e um de fevereiro de dois mil e vinte e dois. .

Conceito: ☐ INS ☐ REG ☐ BOM ☒ EXC

Observações:

_____

_____

DocuSigned by: 364039753882467...
Prof. Dr. Claudomiro de Souza de Sales Junior
ORIENTADOR

DocuSigned by: 1BA3172A5C3C457...
Profª. Drª. Regiane Silva Kawasaki Francês
AVALIADORA

DocuSigned by: 44D0JMA3C2EE477...
Ing. Msc. Ewerton Cristhian Lima de Oliveira
COORIENTADOR

DocuSigned by: 2E86DJC162BC414...
Prof. Dr. Reginaldo Cordeiro dos Santos Filho
AVALIADOR

DocuSigned by: 761CE4B03BF447E...
Prof. Dr. Kauê Santana da Costa
AVALIADOR

DocuSigned by: 2E86DJC162BC414...
Dr. Caio Marcos Flexa Rodrigues
AVALIADOR

To those who refuse to settle for average and strive to attain the impossible.

ACKNOWLEDGMENTS

There are several individuals who accompanied, mentored, or taught me as I grew academically, personally, and professionally. If this manuscript came to be and I am where I am, it is a joint product of their continuous, selfless efforts towards my advancement. I am flattered to receive these contributions, some of which I deferentially recognize here.

Firstly, I would like to thank all the relatives who were with and supported me throughout this journey, from celebrations and reunions to joint endeavors partaken by my aunts, cousins, grandparents, and uncles alike that allowed me to seize important opportunities. A non-exhaustive list of names includes Benedita G. de Souza, Carlos Guilherme L. Moreira, Dalton L. Moreira, Heliana Moreira V. Araújo, Marcelo Antônio S. de Souza, and Nilton Célio G. de Souza. I also pay homage to Léa L. Moreira, my late grandmother. To all of you, my sincerest appreciation for never leaving my side.

I specifically acknowledge my father, Heber L. Moreira, my mother, Iliane Maria G. Souza, and my sister, Paloma Geovanna S. Moreira, for cheering my ups, enduring my downs, and supporting me at every step of the way. My mother has always been a helping shoulder to rely on during bitter moments and the person with whom I like to chat and play the most between my daily chores. My father, a retired professor at UFPa, is the epitome of competence and instruction; a hardworking man with high standards and indisputable competence in anything he sets his mind to do. My sister has also been an everlasting presence in my life, and I am grateful for her companionship and shared moments as we grew and progressed in our lives.

Recently, my father entered his seventies completing law school with distinction and passing the bar examination, effectively adding a third undergraduate degree to his collection and becoming a licensed lawyer. This is the latest of his demonstrations, which always extend beyond words, that he is the eternal scholar I intend to be during my lifetime. By nearing the completion of my first undergraduate degree, I hope to make you, my mother, and my sister just as proud as you make me everyday.

Secondly, I would like to express my recognition for all educators who contributed to my progress. I begin by explicitly thanking Cristina Takae K. Kamizono for her everlasting contributions. During over a decade of work, she showed me the importance of reading and studying, as well as being self-taught, independent, and assiduous. I owe her my passion for our vernacular language, for English, and for my logical and reasoning skills. I am a proud alumnus of the Kumon center she coordinates and completer of the mathematics,

mother-language, and English programs under her guidance.

I also externalize my deference for the professors and academics who, as I pursued my computer science undergraduate degree, went above and beyond with insightful classes, valuable teachings, and impeccable mentoring. In particular, I acknowledge Claudomiro de Souza de S. Júnior for the opportunity he gave me to deepen my knowledge in areas of my interest and for supervising my research work, among which this one. I also mention Caio Marcos F. Rodrigues, Ewerton Christian L. de Oliveira, and Reginaldo Cordeiro dos S. Filho, with whom I had the pleasure of debating on various occasions for hours on end. Additionally, I document here my appreciation to Regiane Silva K. Francês and Josivaldo de S. Araújo not only for the aforementioned reasons, but also for the mentoring on which disciplines and subjects to study, as well as the help with bureaucratic affairs. Lastly, I reiterate my appreciation for all the members of the evaluation board for accepting to dedicate part of their time to appraise this manuscript. Your contributions to my scholarship merit a distinction of their own, and I am privileged to work alongside you.

Thirdly, I would like to acknowledge all the friends I made throughout my journey in this world. Be it in Belém, Coimbra, or elsewhere, I always had the privilege of being accompanied by extremely intelligent, resourceful, and spirited friends. So as not to unwillingly forget anyone, I will restrict myself to only mentioning below the names of those who accompanied me in the pursuit of a bachelor's degree in computer science at UFPa. To the mentioned ones and everyone else, I want you to know how much you mean to me and the lengths I am willing to go to lend you a helping hand. Cheers to all of you.

- Aian Shay Bentes D. Cardoso;
- Antônio Melgacino de S. Neto;
- Arthur Takeshi N. Yoshikawa;
- Eduardo Gil S. Cardoso;
- Felipe de Melo R. e Oliveira;
- Gabriel Afonso P. da Silva;
- Gabriela S. Maximino;
- George Felipe de M. Silva;
- Giovanne César O. de Souza;
- Guilherme Eiji E. Hantani;
- Iury Glabson O. Silva;
- João Lucas D. Silva;
- João Marcelo F. de Almeida;
- João Victor da Silva D. Canavarro;
- Lucas Mesquita R. Ferreira;
- Marco Aurélio Lima do N. Júnior;
- Pedro Victor A. Melo;
- Rafael da S. Lisboa;
- Renan F. Cunha; and
- Vitor N. Cantão.

~ Igor Matheus S. Moreira.

"Again, you can't connect the dots looking forward; you can only connect them looking backwards. So you have to trust that the dots will somehow connect in your future. You have to trust in something—your gut, destiny, life, karma, whatever—because believing that the dots will connect down the road will give you the confidence to follow your heart, even when it leads you off the well-worn path, and that will make all the difference."

(Steve Jobs in a commencement address)

"This whole time I thought changing the world was something you did, an act you performed, something you fought for. I don't know if that's true anymore. What if changing the world was just about being here, by showing up, no matter how many times we get told we don't belong, by staying true even when we're shamed into being false, by believing in ourselves even when we're told we're too different. And if we all held onto that, if we refuse to budge and fall in line, if we stood our ground for long enough, just maybe—the world can't help but change around us."

(Elliot Alderson in Mr. Robot)

Abstract

Oral bioavailability is a desirable property in drug development. Virtual screening of compounds according to their properties with computational intelligence can accelerate the prediction of their human intestinal absorption (HIA). Despite the existence of studies aimed at predicting HIA of compounds, dimensionality reduction (DR) techniques that extract features are seldom employed to enable visual-exploratory analyses and pre-process data for machine learning (ML) algorithms. This work applies six DR projectors (ivis, KPCA, PCA, PCS, TSVD, and UMAP) to produce two- and three-dimensional projections alongside four ML classifiers (KNN, MLP, RF, and SVM) in predicting HIA of small molecules, an effort that encompassed the analysis of fifty-two pipelines. Results demonstrate that, despite reducing the dimensionality by more than 98%, DR-encompassing pipelines still delivered competitive results while also facilitating visualization, demonstrating the viability and potential of DR via feature extraction as an automated pre-processing step.

Keywords: chemoinformatics, computational pharmaceutics, dimensionality reduction, drug discovery and development, feature extraction, human intestinal absorption, machine learning.

Resumo

Biodisponibilidade oral é uma propriedade desejável no desenvolvimento de drogas. A triagem virtual de compostos de acordo com suas propriedades com inteligência computacional pode acelerar a predição de sua absorção intestinal humana (HIA). A despeito da existência de estudos almejando predizer a HIA de compostos, técnicas de redução de dimensionalidade (DR) que extraem características são raramente empregadas para possibilitar análises visual-exploratórias e pré-processar dados para algoritmos de aprendizado de máquina (ML). Este trabalho aplica seis projetores de DR (ivis, KPCA, PCA, PCS, TSVD e UMAP) para produzir projeções bi e tridimensionais conjuntamente com quatro classificadores de ML (KNN, MLP, RF e SVM) na predição de HIA de pequenas moléculas, um esforço que englobou a análise de cinquenta e dois *pipelines*. Os resultados demonstram que, a despeito de reduzir a dimensionalidade em mais de 98%, os pipelines envolvendo DR ainda apresentaram resultados competitivos enquanto também facilitaram a visualização, demonstrando a viabilidade e o potencial de técnicas de DR via extração de características como uma etapa automatizada de pré-processamento.

Palavras-chave: quimioinformática, farmacêutica computacional, redução de dimensionalidade, descoberta e desenvolvimento de drogas, extração de características, absorção intestinal humana, aprendizado de máquina.

LIST OF FIGURES

LIST OF TABLES

LIST OF ABBREVIATIONS AND ACRONYMS

ADME	absorption, distribution, metabolism, and excretion
ADMET	absorption, distribution, metabolism, excretion, and toxicity
AI	artificial intelligence
ANN	Artificial Neural Network
AUC	area under the ROC curve
BO	Bayesian optimization
CV	cross-validation
DR	dimensionality reduction
DS	data science
DT	Decision Tree
Fcsp^3	fraction of sp^3-hybridized carbon atoms
FWE	family-wise error
GI	gastrointestinal
GS	grid search
HBA	hydrogen-bond acceptors
HBD	hydrogen-bond donors
HIA	human intestinal absorption
HP	hyper-parameter
HPO	hyper-parameter optimization
HSVM	Hierarchical SVM
IoT	internet of things
KDD	knowledge discovery in databases

KNN K-Nearest Neighbors

KPCA Kernel PCA

LDA Linear Discriminant Analysis

logP octanol–water partition coefficient

MCC Matthews correlation coefficient

ML machine learning

MLP Multi-Layer Perceptron

MW molecular weight

NRB number of rotatable bonds

PCA Principal Component Analysis

PCS Polygonal Coordinate System

PLS Partial Least Squares

PNN Probabilistic Neural Network

PES Pyramidal Embedding System

RF Random Forest

ROC receiver operating characteristic

SMILES Simplified Molecular Input Line Entry System

SVM Support Vector Machine

tPSA topological polar surface area

t-DNE t-Distributed Deterministic Neighbor Embedding

t-SNE t-Distributed Stochastic Neighbor Embedding

TSVD Truncated Singular Vector Decomposition

UMAP Uniform Manifold Approximation and Projection

CONTENTS

CHAPTER

1

INTRODUCTION

Obtaining insights from real-world phenomena by means of data is demonstrably paramount (HUSSAIN; HUSSAIN, 2007; CASTELLETTI; LOTOV; SONCINI-SESSA, 2010; ŽILINSKAS, 2015). Data science (DS) has been successfully employed to attain those goals in real-time applications (HABEEB et al., 2019), human-computer interaction (DIX, 2017), memory analysis (FU; SONG; ZHU, 2016), and others, ranging from dynamic storage, data mining, machine learning (ML), and statistical analysis to data visualization (CHEN; ZHANG, 2014). These can be seen as applications of the knowledge discovery in databases (KDD) process (a concept introduced by FAYYAD, 1996), as they employ computational theories and tools from the ever-growing pool of available data to extract useful information. The emergence of Big Data builds on this and heralds a new wave of productivity growth (JIN et al., 2015; ALNUAIMI et al., 2019).

Multiple health-related disciplines involve highly multivariate information, e.g., single-cell transcriptomics analysis (BECHT et al., 2019; LINDERMAN et al., 2019; SZUBERT et al., 2019), disease investigation (NAZIR et al., 2020; LIU, Y. et al., 2021), and evidence-based pandemic management (AHMED, I. et al., 2021). One such discipline is pre-clinical drug discovery and development, also termed drug formulation design (DANISHUDDIN et al., 2021; VIJAYAN et al., 2021; BANNIGAN et al., 2021). Drug formulation came to rely on data-intensive, computation-based research patterns to tackle data with ever-increasing scale and complexity. Integrating artificial intelligence (AI) and multi-scale modeling techniques in this context compensates the scientists' limitations when handling Big Data by leveraging fitting and generalization abilities of learning methods, thus significantly saving time and resources (WANG, W. et al., 2021).

One important DS application in the realm of biosciences is computer-based (i.e., *in silico*) pre-clinical compound screening to design oral drugs. According to Youhanna and Lauschke (2021), "oral drug administration constitutes the most convenient route of drug delivery." When devising a formulation strategy for oral drugs, one of the major stages

is the screening of compounds that can be absorbed by the intestinal epithelial barrier (YOUHANNA; LAUSCHKE, 2021). The intestinal epithelium is the key component of the intestinal mucosal barrier, a semipermeable layer that absorbs relevant elements while preventing the entry of harmful ones (FARRÉ et al., 2020; YOUHANNA; LAUSCHKE, 2021). The absorption rate of drug and bioactivity compounds, or human intestinal absorption (HIA) rate (YANG, M. et al., 2018; NAYARISSERI et al., 2021; DANISHUDDIN et al., 2021), is decisive in determining their bioavailability in the circulatory system, which is responsible for delivering them to the molecular targets wherein the devised therapeutic effects operate (STORPIRTIS; AL., 2011; ROSENBAUM, 2017).

ML models have been used as a modality of *in silico* testing to optimize the drug discovery and development process, which traditionally entails multiple resource-intensive, time-consuming *in vitro* and *in vivo* experiments (YANG, M. et al., 2018; BANNIGAN et al., 2021). Predicting HIA is paramount for oral drugs, since their efficacy is particularly susceptible to gastrointestinal (GI) permeability. *In silico* testing, however, has challenges of its own: the GI physiology is complex and might vary depending on health, age, and sex (DAHLGREN; LENNERNÄS, 2019; MADLA et al., 2021). Furthermore, chemical compounds can be described in a multitude of ways (KUMAR et al., 2017; CARRACEDO-REBOREDO et al., 2021), raising the issue of selecting the best descriptors for the problem at hand. This variety of features, which sometimes surpasses the number of available compounds, is one of the characteristics associated with Big Data (LANEY, 2001).

Big Data arguably not only hinder the task of exploiting data in devising ML-based predictive models, but also the one of visualizing it. These obstacles (mostly attributed to the dimensionality curse, as seen in SONG, M. et al., 2013; CHEN; ZHANG, 2014; JIANG; LU; CHOO, 2018; DRYDEN; HODGE, 2018; VIJAYAN et al., 2021) prompted a reassessment on how to pre-process and portray them, paving the way for alternatives such as reducing their dimensionality (KIM; LEE, 2014; SEWELL, 2018). In the realm of ML, dimensionality reduction (DR) techniques are estimators that aim to make high-dimensional data manageable by mapping them into meaningful, lower-dimensional representations or projections (MAATEN; POSTMA; HERIK, 2009; BURGES, 2010; ENGEL; HÜTTENBERGER; HAMANN, 2012; WANG; SUN, 2015; ESPADOTO et al., 2019). These projections can be used to visually explore data with plotting techniques that only support few dimensions and build predictive models out of a simplified feature space, potentially speeding up runtimes or improving the performance of the resulting model.

To facilitate the use of visualization techniques, automatize data pre-processing, and possibly create better-performing predictive models for HIA, this work investigates the pertinence of employing DR techniques in ML pipelines to help optimize oral delivery. This computational pharmaceutics application involving a multivariate data set intends to ascertain whether the application of DR-encompassing ML pipelines for HIA prediction is

advantageous by means of different perspectives and multiple measurements. Hence, the conclusions of this work shall aid in knowing some of the main projectors and classifiers available, in performing the KDD process to virtually screen small molecules, and in conscientiously choosing them in similar Big Data contexts to support the KDD process and extract the so-called Big Knowledge (LU et al., 2019).

The remainder of this chapter is divided as follows: Section 1.1 further contextualizes the problem domain; Section 1.2 explores the knowledge discovery process, challenges brought by Big Data scenarios, and how DR techniques can assuage them; Section 1.3 explores similar work in the literature and how they compare against the proposed approach; Section 1.4 justifies the existence of this work; Section 1.5 specifies the objectives to be achieved herein; and Section 1.6 clarifies how the rest of this manuscript is structured.

1.1 Oral absorption in drug research and development

The pharmaceutical formulation of a new drug begins with the design or research of a bioactive substance and ends with its successful employment in producing a commercial drug product. As described by Bannigan et al. (2021), this process customarily entails combining inert materials and excipients with active pharmaceutical ingredients to produce viable drug products with desired properties, such as enhanced efficacy, longer-acting therapeutic effects, reduced side effects, and extended shelf-life. This is a nontrivial endeavor due to the complexity, costs, and risks involved. Producing a drug can take from ten to fifteen years and almost US$3 billion on average (BANNIGAN et al., 2021; CARRACEDO-REBOREDO et al., 2021). Of the drugs that manage to get into the clinical phase, 80-90% fail (VIJAYAN et al., 2021). Bannigan et al. (2021) remarked that "although the traditional approach to formulation development has delivered successful drug products to patients, it relies on several inherently time-consuming and often inefficient steps."

By considering the precipitous cost of this process and the strict scrutiny of regulatory agencies, pharmaceutical and biotechnology companies are striving to innovate and adopt new technologies to maximize productivity, minimize costs, and ultimately ensure economic sustainability while bringing new drugs to market in a timely manner (BANNIGAN et al., 2021; VIJAYAN et al., 2021). Hence, reliably screening pharmaceutical compounds by means of specific factors (i.e., properties) is crucial in drug design to ensure their proper selection before *in vitro* and clinical trials (KUMAR et al., 2017; CARRACEDO-REBOREDO et al., 2021).

The increase in number of chemical compounds notwithstanding, neither *in vivo* nor *in vitro* methods scale well to handle current demands of the pharmaceutical industry for new drugs (DANISHUDDIN et al., 2021), especially those targeting poorly-treatable or untreatable diseases. Experimentally measuring properties related to pharmacokinetics—

how drugs move in the body—and pharmacodynamics—how the body reacts to drugs—is expensive and time-consuming (SINGH; GUPTA; BASANT, 2015; DANISHUDDIN et al., 2021). Although it might be argued that *in vivo* assays remain a gold standard in pre-clinics as a means to reproduce biological processes mediated by systemic blood circulation, it is difficult do discern and decouple all involved factors in the passage of a molecule across several biological barriers (FEDI et al., 2021). Furthermore, extrapolating results observed in animals to humans is nontrivial due to species differences in transporters, enzyme expressions, and immune responses (DANISHUDDIN et al., 2021; FEDI et al., 2021).

Contrariwise, *in silico* experiments do not rely on living organisms or in artificial, extrinsic environments; instead, it consists of computational simulations of biological processes (CARRACEDO-REBOREDO et al., 2021). Adopting them can be a strategy to mitigate dependency on animal data (DANISHUDDIN et al., 2021). As such, *in silico* experiments have become an interesting, less resource-intensive alternative to effectively screen for potential drugs and bioactive compounds (CARRACEDO-REBOREDO et al., 2021; DANISHUDDIN et al., 2021). One application of this is ML-assisted drug discovery and delivery (ELBADAWI; GAISFORD; BASIT, 2021; VIJAYAN et al., 2021):

> Embodying a system scale and big data driven vision, AI and machine learning algorithms have initiated the revolution within the pharmaceutical industry that has reinvented drug development over the past decade. Game-changing AI technologies have been widely developed and employed throughout the whole drug development pipeline that may involve drug screening, drug reposition, drug design and synthesis, as well as drug clinical trial design and implementation. Drug screening, in particular, is being transformed as AI models have proven to be reliable and efficient in unfolding the biological target protein's structure, disclosing the drug's physicochemical properties and toxicities, and predicting the drug-target interactions. In addition, AI has begun to reshape the key steps of clinical trial, such as patient cohort selection/recruiting and patient monitor *[sic]*, which are the main causes for high trial failure rates and the main stumbling blocks in the drug development pipeline (HE; LEANSE; FENG, 2021).

While regarded as a computer pharmaceutics application, drug discovery and development also closely relates to chemoinformatics, as it involves chemical information retrieval and extraction, computer-aided drug synthesis, and chemical space exploration to model quantitative structure-activity or structure-property relationships and predict how chemical modifications might influence biological behavior (LO et al., 2018).

As antimicrobial resistance gains global momentum and infectious diseases become the leading cause of morbidity and mortality (*vide* the COVID-19 pandemic), employing AI to assist in drug research and development enables scientists to promote optimizations across multiple steps of this process (HE; LEANSE; FENG, 2021). Examples of that are quantitative structure-activity relationship studies targeting HIA (ZHAO et al., 2001; DAHLGREN; LENNERNÄS, 2019).

HIA refers to the permeation ability of a given chemical compound at the intestinal level, a decisive process in its bioavailability (DAHLGREN; LENNERNÄS, 2019; FEDI et al., 2021). Other factors include dissolution, solubility, luminal stability, and intestinal transit time (DAHLGREN; LENNERNÄS, 2019). Oral bioavailability is a major consideration, as oral drug administration is preferred in terms of convenience, ease of use, and patient compliance (WANG, N.-N. et al., 2017; ESAKI et al., 2019; YOUHANNA; LAUSCHKE, 2021). Almost 60% of unique drugs approved by the Food and Drug Administration in 2018 and 90% of commercially available drugs are orally administered (ESAKI et al., 2019; LEE et al., 2020), attesting its preference by pharmaceutical companies. Thus, testing the permeability of chemical compounds through organic membranes is essential in pre-clinical drug development (TOZER; ROWLAN, 2016; FEDI et al., 2021).

Under this context, one problem that can be explored *in silico* with the assistance of AI is ML-assisted HIA prediction of small molecules to optimize oral drug discovery and development. However, given how chemical compounds can be described in a wide spectrum of ways (KUMAR et al., 2017; CARRACEDO-REBOREDO et al., 2021), it is important to consider the impact of high-dimensional data sets on model prediction performance (i.e., the curse of dimensionality customarily regarded as a trait of Big Data) and how to mitigate it by means of DR to enable and facilitate the KDD process.

1.2 Knowledge discovery in Big Data

As originally defined by Fayyad (1996), the knowledge discovery in databases (KDD) field is concerned with making sense of data by means of methods and techniques. Its process entails several steps that fall into one of three groups (D'ALCONZO et al., 2019):

- *Data input*, which involves the management of data from collection in its raw form to delivery in its processed form.
- *Data analysis*, which receives the processed data and extracts information from it (e.g., patterns, relations, rules).
- *Data output*, which turns information into knowledge by means of quality measurements and visualization techniques, ultimately aiding in decision-making processes.

If the definition of the KDD process was to be extended to contemplate Big Data, the terminology discussed by Lu et al. (2019) could be used to characterize it as the conversion of Big Data into Big Knowledge, i.e., "a massive set of structured knowledge elements, where a knowledge element may be a concept, an entity, a datum, a rule or any other computer operable information element." As will be elaborated upon, the change in scale brought about by Big Data impacts every step of KDD (D'ALCONZO et al., 2019).

Albeit regarded as an era-defining concept, Big Data lack a consensus in their characterization (WAMBA et al., 2015; FERNÁNDEZ et al., 2015). Some attempts break the term into a varying set of Vs: an initial definition attributed to Laney (2001) comprises volume, velocity, and variety (WAMBA et al., 2015; ANG; GE; SENG, 2020; XU et al., 2017; SONG, J. et al., 2020), but recent definitions include value, variability, veracity, volatility, and visualization (FERNÁNDEZ et al., 2015; JIN et al., 2015; SIVARAJAH et al., 2017; MEHMOOD; ANEES, 2020). Wamba et al. (2015) summarize it as a holistic approach to manage, process, and analyze these Vs to obtain insights, deliver value, measure performance, and establish competitive advantages.

Big Data contexts such as real-time applications and internet of things (IoT) devices brought unprecedented awareness to data analysis techniques in research and development (YAQOOB et al., 2016; AHMED, E. et al., 2017; HABEEB et al., 2019; REHMAN et al., 2019). While this profusion of knowledge and ubiquitous data collection amounted to breakthroughs in scientific disciplines such as health and science, it also brought challenges: Big Data accentuated bottlenecks of popular data analysis methods, thus impelling extraordinary measures and new tools to make use of massive knowledge in a limited time period (CHEN; ZHANG, 2014; JIN et al., 2015; YAQOOB et al., 2016; GARDINER et al., 2018; HABEEB et al., 2019). Taking IoT applications as an illustration, potential problems include modeling complexity due to large-scale processes, data redundancy introduced by multidimensional data asynchronously collected across distributed nodes, and algorithmic scalability (LI, F. et al., 2020). Such hindrances are also becoming commonplace in drug formulation, as chemical Big Data is enabling novel ML-based applications (DANISHUDDIN et al., 2021; PATEL; SHAH, 2021).

This work focuses on two data analysis tasks that are hindered by massive, high-dimensional knowledge and can be facilitated with DR. They interact with the *variety* aspect of Big Data (i.e., multi-sourced, multi-dimensional information) and are, namely:

- *Data inspection for knowledge discovery*, which entails *visualization* (data presentation in an intelligible manner).
- *Data exploitation for predictive purposes*, which generates *value* (conversion of data in relevant insights).

Visualization is a *sine qua non* component of exploration, the first phase of data analysis in which one makes sense of data prior to more focused procedures (ENGEL; HÜTTENBERGER; HAMANN, 2012). It is also used in the KDD process as a means of interpreting extracted patterns (FAYYAD, 1996). According to Hussain and Hussain (2007), visual data analysis (or scientific visualization) has aided the development of many spheres of knowledge. It copes with an array of approaches to enable visual information-extraction processes (SEWELL, 2018), many of which targeted at multidimensional

information (PATHAK; PATHAK, 2020). Nevertheless, visualizing high-dimensional data is regarded as a complex, cumbersome problem: in addition to their structural uncertainties, intuitively exposing multidimensional knowledge with traditional methods is nontrivial due to lackluster functionality, poor scalability, and protracted execution (MAATEN, 2014; CHEN; ZHANG, 2014; SEWELL, 2018). Most conventional approaches (e.g., scatter plots) do not support the depiction of more than three dimensions, and the ones that do end up merely squeezing more dimensions in the same view, yielding non-intuitive, limited visualizations (PATHAK; PATHAK, 2020).

A natural step after exploratory analysis is exploitation for predictive purposes. One means to do so is via ML, a branch of AI. ML estimators are capable of automatically interpreting data sets to predict the output of unforeseen instances (HE; LEANSE; FENG, 2021; CARRACEDO-REBOREDO et al., 2021). ML applications can be observed from fields as varied as crime management (FENG et al., 2019), malware detection (GUPTA; RANI, 2020), self-healing 5G networks (OMAR; KETSEOGLOU; NAFFAA, 2021), and drug-microbiome interaction prediction (MCCOUBREY et al., 2022).

Conventional ML techniques such as K-Nearest Neighbors (KNN), Decision Tree (DT), and Support Vector Machine (SVM) are appealing by virtue of their relative simplicity, interpretability, and predictive performance, thus facilitating their employment by researchers outside the computer science realm (ELBADAWI; GAISFORD; BASIT, 2021). In contrast, they might underperform when the dimensionality of the data is high (ELBADAWI; GAISFORD; BASIT, 2021) or the relevant hyper-parameters (HPs) are not properly optimized (YANG; SHAMI, 2020). The possibility of employing more powerful models and hyper-parameter optimization (HPO) notwithstanding, the diversity, scale, and complexity of Big Data prompt novel approaches to enable their adoption in real-world applications (LO et al., 2018; NAZIR et al., 2020):

> A common ML presumption is that algorithms can learn better with more data and consequently provide more accurate results. However, massive datasets impose a variety of challenges because traditional algorithms were not designed to meet such requirements. For example, several ML algorithms were designed for smaller datasets, with the assumption that the entire dataset can fit in memory. Another assumption is that the entire dataset is available for processing at the time of training. Big Data break these assumptions, rendering traditional algorithms unusable or greatly impeding their performance (L'HEUREUX et al., 2017).

It can be argued that the actual challenge of Big Data does not lie so much in collecting it as it does for managing and interpreting it, or rather, in extracting Big Knowledge out of it (LV et al., 2017). Despite the applicability of the KDD process to any data analytics, the Vs of Big Data impose fundamental challenges to the methodologies behind each step (D'ALCONZO et al., 2019). DR techniques can alleviate some of these symptoms, expediting and even enabling otherwise impracticable activities.

DR is a broadly employed tool in statistics and ML that aims at representing high-dimensional data in low-dimensional spaces while preserving relevant structural properties (ESPADOTO et al., 2019; FLEXA; GOMES; MOREIRA; ALVES, et al., 2021). Despite being a means of lossy-compressing data, DR can improve the accuracy of ML algorithms by producing a representation that better represents data with less redundancy and attenuated multicollinearity (FLEXA; GOMES; MOREIRA; ALVES, et al., 2021; HOUARI et al., 2016a). Additionally, they can serve as a means to visually grasp causal data relationships, screen for relevant patterns or properties, and determine whether a given data set is clusterable (CHAFFI; TAFRESHI, 2019; HOLIDAY et al., 2019; LI; LI; ZHANG, 2019; FLEXA; GOMES; MOREIRA; ALVES, et al., 2021).

Considering the aforementioned qualities of DR techniques and their relevance in chemoinformatics contexts (MAHMUD et al., 2021; MOZAFARI et al., 2022), they can be seen as a viable alternative to enable visual-exploratory analyses and potentially improve ML-assisted HIA prediction performance of small molecules to optimize oral delivery in drug research and development.

1.3 Related work

The results of the performed literary review are displayed below. This review looked for papers that, akin to this work, trained ML models and pipelines to predict HIA of small molecules in the context of drug research and development. Information on the size and dimensionality of the data sets and the classification algorithms used are exposed, as well as the presence, type, and purpose of DR techniques that were eventually applied.

Aiming at the creation of both quantitative and qualitative (i.e., regression and classification, respectively) permeability prediction models, Singh, Gupta, and Basant (2015) took advantage of the Gradient Boosting algorithm and fitted it to a data set composed of 684 organic molecules. Each molecule is described by means of 211 descriptors. Cross-validation was performed to fine-tune relevant classifier HPs.

To label small molecules in terms of GI absorption and brain penetration, Daina and Zoete (2016) resorted to two molecular descriptors and calculated an elliptical region in the two-dimensional space that included the most well-absorbed compounds while excluding as many poorly absorbed ones. This method is denominated brain or intestinal estimated permeation method, or BOILED-Egg in short.

Ning-Ning Wang et al. (2017) quantified 970 compounds with nine different features types, namely descriptors (two- and three-dimensional), molecular fingerprints (ECFP2, ECFP4, ECFP6, and FP2), and structural fragments (FP4, Estate, and MACCS). The data was used to create a Random Forest (RF) model with fine-tuned HPs.

By means of feature selection, Kumar et al. (2017) reduced the dimensionality from 1,529 to ten molecular descriptors. The chosen features were collected for 1,242 compounds and given as input for Artificial Neural Network (ANN), KNN, Linear Discriminant Analysis (LDA), Partial Least Squares (PLS), and Probabilistic Neural Network (PNN) classifiers to predict whether a given chemical compound is well-absorbed by the human intestine. Shin et al. (2018) also resorted to an ANN to predict the absorption potential of chemicals, with 663 compounds described in terms of 209 features as input data.

Esaki et al. (2019) created a three-class HIA prediction model by using RF, Linear SVM, and Radial SVM. Feature selection was applied via the `Boruta` package in the `R` programming language to select some of the 7,908 collected descriptors, resulting in feature sets ranging from less than thirty to more than five hundred descriptors depending on the employed ML algorithm. Principal Component Analysis (PCA) was used to visually inspect the distribution of the 946 compounds against approved drug data.

To predict positive, neutral, or negative food-provoked changes to drug absorption (also known as food effects), Gatarić and Parojčić (2020) devised a three-step hybrid approach termed hierarchical clustering on principal components. Firstly, DR was performed with PCA, a feature extraction technique, followed by hierarchical clustering to identify the number of clusters. Lastly, several iterations of the K-means algorithm were applied. Eleven physicochemical, biopharmaceutical, and pharmacokinetic properties were selected to describe fifty-three drugs by means of four feature compositions. Likewise, Gavins et al. (2022) selected twenty-three physicochemical properties to describe 311 drugs. The resulting data set was used to train RF, Logistic Regression, SVM, and KNN models. An exploratory data analysis was also performed using t-Distributed Stochastic Neighbor Embedding (t-SNE) to enable the depiction of the data set in two dimensions.

Lee et al. (2020) chose six descriptors to describe sixty-six molecules. The resulting data set served to train a SVM variation termed Hierarchical SVM (HSVM) and generate a nonlinear structure-activity relationship model tying the selected descriptors to intestinal permeability. Similarly, HSVM was used by Ta et al. (2021) to "(...) depict the exceedingly confounding passive diffusion and transporter-mediated active transport". Over a hundred descriptors were enumerated, filtered, and selected via Recursive Feature Elimination.

To foresee observed influx and efflux permeability (i.e., permeability from the apical side to the basal one and *vice-versa*, respectively) of chemicals through the intestinal barrier, Kamiya et al. (2021) selected 196 descriptors to describe 219 chemicals. Univariate, bivariate, and trivariate linear regressions were used, as well as light gradient boosting.

To identify new relationships between approved drugs, therapeutic targets, and diseases to repurpose them, Madugula et al. (2021) performed unsupervised learning with K-Means on 1671 approved drugs. As a pre-processing step, PCA was employed alongside other pre-processing techniques from 6,494 to 1079 features.

In contrast with this scenario of spreading use of classifiers, there is a lack of articles that explore applications that jointly take advantage of DR by means of feature extraction and ML under the context of HIA. From the aforementioned articles, only Gatarić and Parojčić (2020) and Madugula et al. (2021) used feature extraction to perform DR as part of a classification pipeline, and only used PCA at that. No analysis on the produced projections were performed. Moreover, these manuscripts analyzed problems that relate to HIA, and not HIA *per se*. The articles that were directly related to HIA prediction and applied DR did so with feature selection techniques. In terms of visual-exploratory analyses, Esaki et al. (2019) used PCA and Gavins et al. (2022) used t-SNE, but the extracted projections were not used for model training.

1.4 Justification

This work is justified by virtue of its employment of feature extraction for visual inspection and ML-assisted HIA prediction. The literature of the area seldom employs DR via feature extraction methods, and the few articles found that use said techniques only do so as an analysis tool, falling short of exploring its use alongside ML algorithms.

By extensively and systematically scrutinizing different combinations of DR and ML techniques under the context of drug research and development, this work contributes not only to the pharmaceutical, but also to the chemical and computational realms. The discussion contained herein should suffice to guide specialists in picking a competent set of techniques for ML-assisted HIA prediction, ultimately disseminating knowledge about DR techniques that extract features and the possibilities they bring.

To a lesser extent, the findings of this work can also be used by specialists of other areas of chemistry, pharmacy and computing, for this work touches in pervasive subjects in these disciplines (e.g., Big Data and AI) and presents a methodology to handle issues originated by high-dimensional data with DR via feature extraction. Hopefully, specialists of multiple disciplines of knowledge will be able to get acquainted with these techniques and resort to them in their projects as a consequence of this manuscript.

1.5 Objectives

The overarching goal of this manuscript is to ponder on the pertinence of reducing the dimensionality of molecular data by means of feature extraction techniques, aiming at replacing feature selection, enabling exploratory visual analyses, and potentially improving predictive models of HIA. This encompasses the following specific goals:

- Devise a common pipeline structure that receives the data set as input and returns

labels as output, such that only the projectors (if applicable) and classifiers should change between pipelines.

- Choose a representative set of projectors (Ivis, KPCA, PCA, PCS, TSVD, and UMAP) and classifiers (KNN, MLP, RF, and SVM).
- Jointly tweak the HPs of the estimators of each pipeline via Bayesian optimization (BO) with cross-validation (CV) to maximize the mean accuracy of the pipelines.
- Scrutinize the results of the pipeline tuning phase in terms of mean train/validation accuracy and runtime measurements obtained from 10-fold CV.
- Determine the statistical significance of the CV accuracy measurements with Friedman's test, as well as pinpoint affected pipeline pairs with the Conover and Nemenyi *post-hoc* tests whose results are adjusted, if applicable, by the Bonferroni and Hommel p-value adjustment techniques.
- Perform independent testing of the pipelines with a hold-out set of small molecules after applying the best settings found in the HPO process.
- Analyse the independent testing results by visually inspecting the produced projections and inspecting the performance measurements of the trained models in terms of HIA prediction.
- Discuss the pertinence of DR-encompassing ML pipelines, extrapolating the conclusions in hopes of foreseeing potentialities for HIA and related Big Data contexts.

1.6 Manuscript structure

The remaining chapters of this work are organized as follows: Chapter 2 elaborates both on the problem domain and on the selection of projectors and classifiers to be studied herein; Chapter 3 clarifies particularities of the data set to be used, the devised pipeline structure, and the process to fine-tune, test, and compare said pipelines from multiple combinations of DR and ML algorithms; Chapter 4 exposes and discusses the obtained cross-validation and independent testing results; and Chapter 5 concludes this manuscript by summarizing its discoveries and enumerating potential future work.

CHAPTER

2

THEORETICAL BACKGROUND

After exposing the overarching context in which this work is inserted in Chapter 1, this chapter will elaborate on some key topics to further contextualize both the problem domain and the proposed methods. This chapter will conceptualize the HIA process (Section 2.1), molecular descriptors and fingerprints (Section 2.2), DR techniques and ML classifiers (Section 2.3), HPO strategies coupled with CV (Section 2.4), and statistical inference via hypothesis testing (Section 2.5).

2.1 Human intestinal absorption process

As exposed in the previous chapter, oral drug treatment is the preferred administration route for acting drugs by virtue of patient preference, cost-efficient manufacturing, and non-invasiveness (YOUHANNA; LAUSCHKE, 2021; GAVINS et al., 2022), meaning that HIA is a determining factor to be considered (DAINA; ZOETE, 2016). HIA is a pharmacokinetic process based on which a certain amount of a given compound manages to travel from external sources (e.g., oral, dermal, respiratory) through biological membranes, effectively entering into a living organism (CARRASCO-CORREA et al., 2021). A key part of this process is the absorption of the compound by the epithelial barrier of the GI tract (YOUHANNA; LAUSCHKE, 2021), which tends to prevent the absorption and translocation of potentially harmful luminal constituents into the central circulation while allowing the absorption of nutrients and water (TOZER; ROWLAN, 2016). This process can be influenced by intrinsic and extrinsic factors, ranging from age and sex to diet and food intake (GAVINS et al., 2022). Hence, adopting the oral administration route incurs several obstacles for a drug treatment to be successful.

Assessing oral drug bioavailability for novel medicines comprises complex underlying processes including but not limited to solubility in gastrointestinal fluids and permeability in the intestinal membrane (YOUHANNA; LAUSCHKE, 2021). Both must

be proven throughout the drug development process and are influenced by physiological and physicochemical variables. This is particularly true for new chemical entities, since their physicochemical properties are often employed as an indicator of their drug-likeness.

According to Wagner (1981), pharmacokinetics aims to "study the time course of drug and metabolite concentrations or amounts in biological fluids, tissues and excreta, and also of pharmacological response, and to construct suitable models to interpret such data." Four processes that involve several organs (e.g., intestine, liver, kidneys) determine the pharmacokinetics of a drug in the human body: absorption, distribution, metabolism, and excretion (ADME) (STORPIRTIS; AL., 2011; FEDI et al., 2021). Toxicity is also considered, resulting in the absorption, distribution, metabolism, excretion, and toxicity (ADMET) acronym (CARRACEDO-REBOREDO et al., 2021; VIJAYAN et al., 2021). To understand the pharmacokinetic properties of a compound, it is crucial to estimate the efficacy of its intestinal absorption and bioavailability in the circulatory system. This is notably necessary for orally-ingested compounds, as they are subject to all ADME, especially in terms of absorption (FEDI et al., 2021).

When entering the stomach after going from the mouth through the esophagus, the drug encounters an acidic environment and specific enzymes involved in food digestion. Muscular contractions eventually break the drug capsule or tablet into smaller particles,

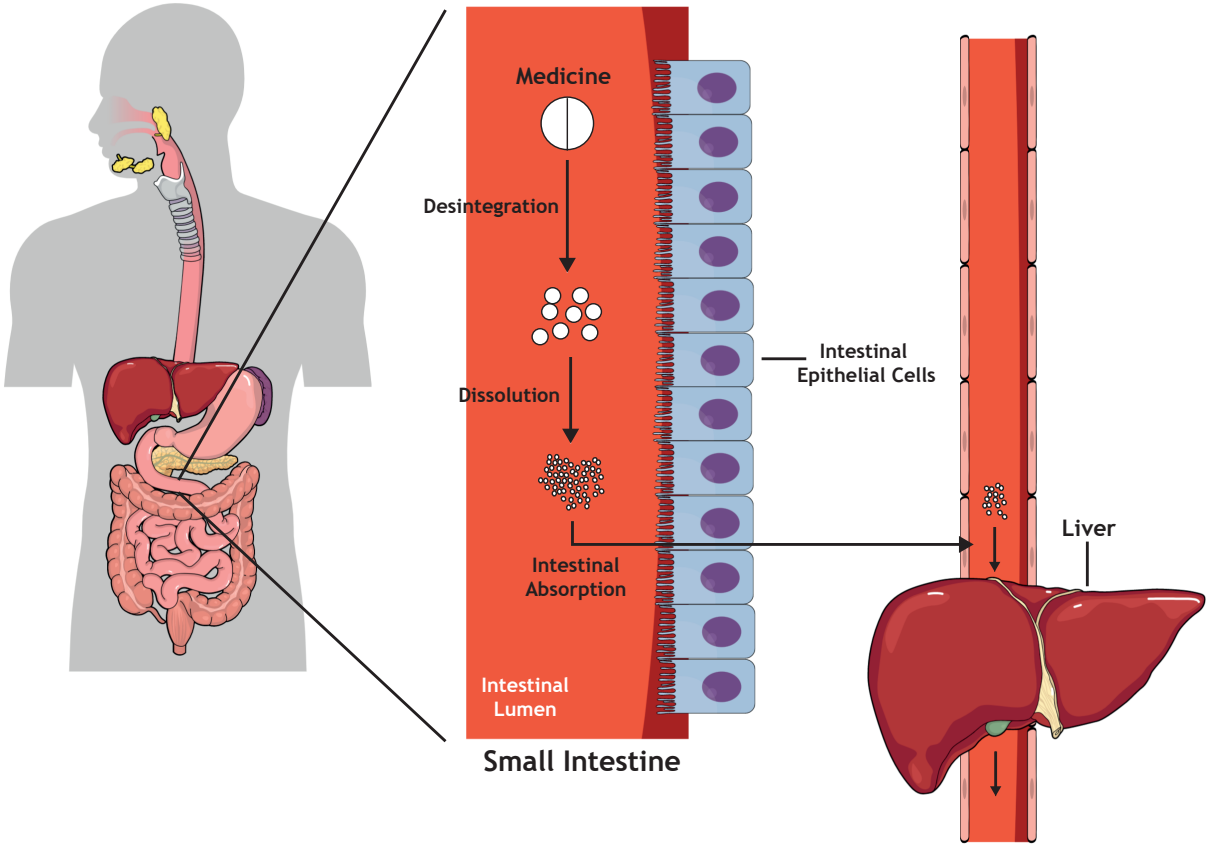


Figure 1 – Main steps of the pharmacokinetics process of HIA. Source: the author.

which can be dissolved in GI fluid. The lipophilicity, hydrophilicity, and solubility of these compounds in the intestinal fluids are crucial factors in this initial part of the absorption process to provide a fast dissolution rate and confer permeability to cross the apical GI membrane (FEDI et al., 2021).

Figure 1 illustrates the process through which a drug undergoes in the small intestine once the GI content reaches this organ. This organ has a columnar epithelium with a significant surface area, an attribute that collaborates with the high absorption of nutrients, water, electrolytes and xenobiotics observed (FEDI et al., 2021).

After the molecules are transported across the outer lipophilic cell membrane—which can be resistant to hydrophilic molecules or drugs with high molecular weights (DAHLGREN, 2018)—, the drug molecules are considered absorbed. Finally, before its introduction into the central blood circulation, an absorbed drug molecule passes through the liver, where it is either metabolized and might have its pharmacological effects altered, or excreted back into the intestines with bile (YOUHANNA; LAUSCHKE, 2021). At the end of this process, the compound is deemed bioavailable, i.e., absorbed and distributed to therapeutic targets (RODRIGUES; FAILLA, 2021).

Predicting whether a given drug would be well-absorbed by the intestines is a task that can be aided by ML. As a prior step to enable this, chemicals must be described in quantitative and categorical terms. This is where molecular descriptors come in.

2.2 Molecular descriptors and fingerprints

Molecular descriptors can be characterized as numerical representations of a given molecule in terms of physicochemical or structural information—or at least, the aspects that can be numerically measured (CARRACEDO-REBOREDO et al., 2021). They can be derived from multiple molecular information sources, including chemical formulas, molecular structures, and interactions with other molecules (VINOTHA; SUNDAR, 2019).

Molecular descriptors play an essential role in chemistry, pharmaceutics, and other fields (VINOTHA; SUNDAR, 2019). In drug research and development, the chemoinformatics task of numerically describing small molecules and exploiting the similarity principle has yielded paramount contributions (FERNÁNDEZ-TORRAS et al., 2022). Molecular descriptors are also used for virtual screening tasks to scan multiple libraries of chemical structures looking for molecules with desirable chemical structure or utility, thus accelerating drug development processes (VINOTHA; SUNDAR, 2019).

There are more than three thousand molecular descriptors, which can be stratified considering the dimensions of the molecular characteristics they describe (KUMAR et al., 2017; CARRACEDO-REBOREDO et al., 2021). They range from simple molecular

properties to intricate three-dimensional and complex molecular fingerprint formulations defined in vectors that span from hundreds to thousands of elements (CARRACEDO-REBOREDO et al., 2021). The task of describing compounds by descriptors and fingerprints for ML-assisted prediction purposes can be regarded as a chemoinformatics process:

> Converting a compound structure into chemical information applicable for machine learning tasks requires multilayer computational processing from chemical graph retrieval, descriptor generation, fingerprint construction to similarity analysis, in which each layer is built upon the successful development of previous layers and often has a substantial impact on the quality of the chemical data for machine learning (LO et al., 2018).

Some aspects must be pondered when choosing molecular descriptors and fingerprints, viz., the biological activity under scrutiny, predictability, and computational costs. Failure to do so might result in sub-optimal feature sets and, ultimately, underperforming predictive models (SANTANA et al., 2021).

In terms of physicochemical properties, some studies related to oral bioavailability of chemicals guide the process of drug discovery and development (BICKERTON; AL., 2012). These studies involve the following properties:

- *Fraction of sp^3-hybridized carbon atoms ($Fcsp^3$)*: the ratio of sp^3 carbons to all carbons of the compound.
- *Hydrogen-bond acceptors (HBA)*: number of bonds where the compound is not covalently attached to the hydrogen.
- *Hydrogen-bond donors (HBD)*: number of bonds where the compound is covalently attached to the hydrogen.
- *Octanol–water partition coefficient (logP)*: the ratio of concentration rate of the compound in the octanol phase to the aqueous phase.
- *Molecular weight (MW)*: the mass of the compound.
- *Number of rotatable bonds (NRB)*: the number of bonds of the compound that allow free rotation around themselves.
- *Topological polar surface area (tPSA)*: the sum of surfaces of polar atoms of the compound.

An assessment for drug-likeness of compounds using physicochemical properties is Linpiski's rule of five (ZHAO et al., 2001). Lipinski et al. (1997) stated that poor absorption is more likely when MW $>$ 500 g/mol, logP $>$ 5, HBD $>$ 5, and HBA $>$ 10. Posteriorly, Veber and al. (2002), proposed additional rules involving NRB and tPSA, which state

that, for good absorption, NRB $\leq$ 10 and tPSA $\leq$ 140 Å. Then, Lovering, Bikker, and Humblet (2009) added that oral drugs should have an average value for $Fcsp^3$ of 0.43.

Lipinski's rules have its critics, since approximately 7.86% of approved oral drugs up to 2016 fails in two or more conditions (BENET; AL., 2016). This might be a consequence of activity cliffs, i.e., compounds with minor differences in functional groups that result in substantially discrepant performance (LO et al., 2018).

In summary, molecular descriptors are fundamental to understand aspects of compounds related to a multitude of biological activities. For drug research and development, permeability and absorption are essential functions that need to be understood to properly evaluate the drug-likeness of compounds. In this context, ML can accelerate the discovery of quantitative structure-activity and structure-property relationships tying molecular descriptors with permeation capacity in the human GI tract.

2.3 Machine learning for projection and classification

As defined in the previous chapter, ML is a branch of AI responsible for developing algorithms and routines that learn from data to perform a specific task (CARRACEDO-REBOREDO et al., 2021). In more formal terms, ML happens when a computer is set to be a learning agent: it observes some data, builds a model based on it, and yields a predictive system that can be used both as a hypothesis about the world and as a piece of software that can solve problems (RUSSELL; NORVIG, 2021, p. 669). Hence, ML algorithms can be defined as *induction* techniques, as they aid in devising general rules from a specific set of observations. This learning process can be performed by four approaches (RUSSELL; NORVIG, 2021, p. 671, 723):

- *Supervised learning*, when the algorithm observes input-output pairs aiming at learning a mapping that properly ties them.
- *Semi-supervised learning*, when the algorithm resort to a few labels to mine more information from a larger collection of unlabeled samples.
- *Unsupervised learning*, when the algorithm learns patterns based exclusively on the input, without the support of output labels.
- *Reinforcement learning*, when the algorithm evolves its behavior based on a system of rewards and punishments.

Although ML is customarily associated with tasks such as classification, clustering, and regression, DR also lies under this umbrella, as it essentially learns a mapping from high-dimensional vectorial spaces to representations composed of fewer dimensions

(MAATEN; POSTMA; HERIK, 2009; BURGES, 2010; ENGEL; HÜTTENBERGER; HAMANN, 2012; WANG; SUN, 2015; ESPADOTO et al., 2019).

Ultimately, the goal of employing ML algorithms in a pipeline is to generate a model that competently devises a hypothesis from the data it processed. A series of factors impact how reasonable the resulting hypothesis is, including the HP settings and algorithms that were chosen. Poor choices in these matters might result in one of two things (RUSSELL; NORVIG, 2021, p. 673):

- Models with *high bias*, i.e., that have a restricted hypothesis space and are likely to *underfit*, meaning that they will not be able to capture patterns of the problem conveyed by the data.
- Models with *high variance*, i.e., that substantially change the hypothesis due to fluctuations contained in the data and are likely to *overfit*, meaning that they learn patterns pertaining to the training data *per se* instead of only capturing underlying patterns related to the conveyed problem.

Therefore, by considering these concepts when devising predictive models, one is considering the often present bias-variation trade-off:

> (...) a choice between more complex, low-bias hypotheses that fit the training data well and simpler, low-variance hypotheses that may generalize better. Albert Einstein said in 1933, "the supreme goal of all theory is to make the irreducible basic elements as simple and as few as possible without having to surrender the adequate representation of a single datum of experience." In other words, Einstein recommends choosing the simplest hypothesis that matches the data. This principle can be traced further back to the 14th-century English philosopher William of Ockham. His principle that "plurality [of entities] should not be posited without necessity" is called Ockham's razor because it is used to "shave off" dubious explanations.
>
> Defining simplicity is not easy. It seems clear that a polynomial with only two parameters is simpler than one with thirteen parameters. (...) However, (...) deep neural network models can often generalize quite well, even though they are very complex—some of them have billions of parameters. So the number of parameters by itself is not a good measure of a model's fitness. Perhaps we should be aiming for "appropriateness," not "simplicity" in a model class (RUSSELL; NORVIG, 2021, p. 673).

As empirically attested by Fernandez-Delgado et al. (2014) for 179 classifiers arising from nineteen families and 121 data sets, the no-free-lunch theorem[1] (WOLPERT, 1996; WOLPERT; MACREADY, 1997) holds true for classification tasks. In other words: there is no one-size-fits-all algorithm; instead, one should conscientiously choose which algorithms to use based on the idiosyncrasies of the problem domain and the data under scrutiny

[1] Please refer to the no-free-lunch.org website for a holistic overview on no-free-lunch theorems.

(SCHAFFER, 1993). This definition also applies to HPs, as different settings suit different problems (YANG; SHAMI, 2020). This will be elaborated upon on Section 2.4.

Herein, Section 2.3.1 describes DR algorithms (also referred henceforth as *projectors*), whereas Section 2.3.2 characterizes classification techniques (also referred henceforth as *classifiers*), both equally important for the pipelines proposed in this work. These sections also define the selection of projectors and classifiers used herein.

2.3.1 Projectors

Dimensionality reduction (DR) techniques map data into a meaningful, lower-dimensional representation or projection (MAATEN; POSTMA; HERIK, 2009; BURGES, 2010; ENGEL; HÜTTENBERGER; HAMANN, 2012; WANG; SUN, 2015; ESPADOTO et al., 2019). Quoting Anowar, Sadaoui, and Selim (2021), DR techniques "aim to address the curse of dimensionality that makes machine learning algorithms incompetent". DR is employed in statistics, ML, and information theory to obtain a representative feature space that, by being more sensible to the problem domain, best describes data and preserves important structural properties, viz., dissimilarities, cluster shapes, probability distributions, and neighboring relationships (ROWEIS; SAUL, 2000; HOUARI et al., 2016b; GODDARD et al., 2018; KIM, 2018; FLEXA; SANTOS, et al., 2019). It is an important data mining tool, maximizing model accuracy and trust in decision processes (SONG, M. et al., 2013; HOUARI et al., 2016b; OLSON; JUDD; NICHOLS, 2018; LI; LI; ZHANG, 2019). Its applications range from filtering, compression, regression, classification, and analysis to visualization (ENGEL; HÜTTENBERGER; HAMANN, 2012).

One can perform DR by selecting or extracting features (NAHID; KONG, 2017; NWEKE et al., 2018; LIU, Z. et al., 2019; AYESHA; HANIF; TALIB, 2020). Techniques on the former category elect a subset of input variables or predictive characteristics. It frequently happens by filtering features agnostically to the ML algorithm employed (LI, Y. et al., 2012; WAN et al., 2016) or by wrapping them—that is, selecting features to maximize the accuracy of the chosen ML algorithm (PEREIRA; PLASTINO, et al., 2018; BRUNELLO et al., 2019). Selecting features is arguably the easiest form of DR; however, despite allowing some degree of perception on the data at hand, the subspace combinations to be analyzed grow quadratically as dimensionality increases. Moreover, it is virtually assured that spatial within-data arrangements will be partly lost or distorted in the process due to an inferior resulting data dimensionality when compared to the intrinsic one—the smallest variable set capable of integrally retaining data-inherent structures (HOULE, 2017; HOULE; SCHUBERT; ZIMEK, 2018). Feature extraction (also termed feature transformation) techniques circumvent this by transforming (i.e., embedding or mapping) high-dimensional spaces into lower-dimensional ones (AYESHA; HANIF; TALIB, 2020).

DR via feature extraction is carried out by applying linear or non-linear trans-

Table 1 – Selected DR techniques stratified by transformation type and ordered by ascending debut date. Source: the author.

	Property	Learning	Debut paper
Transformation	Projector		
linear	PCA	unsupervised	Pearson (1901)
	TSVD	unsupervised	Hanson (1971)
	PCS	unsupervised	Flexa, Gomes, Viademonte, et al. (2019)
non-linear	KPCA	unsupervised	Schölkopf, Smola, and Müller (1997)
	UMAP	both	McInnes, Healy, and Melville (2018)
	Ivis	both	Szubert et al. (2019)

formations. Although supervised variants are available, the projection process is usually non-supervised. Linear *modi operandi* choose a linear projection that befits the data. Cunningham and Ghahramani (2015) define linear DR as a cornerstone of high-dimensional data analyses by virtue of its "simple geometric interpretations and typically attractive computational properties". Although they maximize the originally seen variation in the transformed data, they usually do not preserve complex structures (OLSON; JUDD; NICHOLS, 2018; LI; LI; ZHANG, 2019; ZHOU; WEISKOPF, 2019).

Conversely, non-linear feature extraction assumes the dimensionality of data to be artificially high (i.e., relevant data lies in a manifold within the high-dimensional space) and can be thought of as a means to generalize non-linear structures (HALKO; MARTINSSON; TROPP, 2009; KIM; LEE, 2012; ORSENIGO; VERCELLIS, 2013; OLSON; JUDD; NICHOLS, 2018; TURAGA; ANIRUDH; CHELLAPPA, 2020). It is suitable in situations where it is paramount to keep low-dimensional representations of similar objects closer, which is seldom possible with linear embeddings (MAATEN; HINTON, 2008). Non-linear techniques are deemed superior in describing the distribution of complex data that discard the linearity assumption while preserving representative structures (TIAN; TAO, 2020)—as is the case for quantitative structure-activity and structure-property relationship modeling (LO et al., 2018). However, they tend to be computationally costlier (MAATEN; POSTMA; HERIK, 2009; KIM; LEE, 2012).

It is worth mentioning that, in contrast with DR based on feature selection, analysis of features that were extracted by DR techniques are limited for they are compressed, transformed representations of the original information. Fortunately, some techniques have support for inversely mapping instances from the projected feature space to the original one, thus helping alleviate the issue.

A wide spectrum of DR methods has been proposed in the last decades (WANG; SUN, 2015; HOUARI et al., 2016b; SHARIFZADEH et al., 2017; KIM, 2018; ESPADOTO et al., 2019; AYESHA; HANIF; TALIB, 2020). As seen in Table 1, which describes the

chosen projectors in terms of transformation and learning types, a representative selection of approaches was made for this work. This was made considering their ability to produce at least two- and three-dimensional representations; their capability to produce trained models for subsequent transformation of unforeseen data; and their popularity or novelty. Three linear and three non-linear algorithms were chosen. The linear projectors are PCA (Section 2.3.1.1), TSVD (Section 2.3.1.2), and PCS (Section 2.3.1.3). The non-linear ones are KPCA (Section 2.3.1.4), UMAP (Section 2.3.1.5), and Ivis (Section 2.3.1.6).

2.3.1.1 Principal Component Analysis

One of the most popular approaches to DR is the direct transformation of data: axis transformations targeting lower-dimensional representations with optimal variability preservation. Principal Component Analysis (PCA), introduced by Pearson (1901), is arguably the best known and most popular linear DR technique of this kind (ZHANG; DUBAY; CHAREST, 2015; ZHONG; ENKE, 2017; SZUBERT et al., 2019). Its routine involves identifying the orthogonal basis (i.e., the loading matrix) that maximizes the intra-variability of a set of points or vectors. In more specific terms, PCA is an algebraic, HP-free procedure that orthogonally transforms a group of possibly correlated variables into a set of linearly non-correlated principal components (JOLLIFFE, 2002). The canonical PCA only supports unsupervised learning.

Some of the fields wherein PCA applications can be encountered are drug discovery and biomedical data (GIULIANI, 2017), structural damage identification (PATHIRAGE et al., 2018), and diabetes mellitus prediction (ZOU et al., 2018).

2.3.1.2 Truncated Singular Vector Decomposition

Another linear DR technique, Truncated Singular Vector Decomposition (TSVD) can be defined as a similar technique to PCA in the sense that both are a means of performing low-rank matrix approximation (HALKO; MARTINSSON; TROPP, 2009). Similarly to PCA, the canonical TSVD only supports unsupervised learning.

TSVD is an established algorithm whose popularity arose by virtue of its applicability to ill-posed, interdisciplinary problems (XU, 1998). One advantage of TSVD lies in its ability to efficiently handle sparse matrices (ANOWAR; SADAOUI; SELIM, 2021).

2.3.1.3 Polygonal Coordinate System

Polygonal Coordinate System (PCS) is a linear, geometric DR approach proposed by Flexa, Gomes, Viademonte, et al. (2019) that tends to surpass previous techniques in terms of asymptotic time/space complexity and global structure preservation. It is a hyperparameter-free routine whose distinction is the construction of a polygonal interspace

with the data set features, by means of which high-dimensional instances are mapped into lower-dimensional representations.

The PCS routine is composed of three steps: data normalization (where data is normalized under a single range), interspace construction (where the feature vectors of the original data set are used as edges of a polygonal interspace via rotation and translation operations), and 2D/3D data mapping (where the mapping is produced by averaging the edges of the interspace). This approach has linear-time complexity and supports batch-loading, thus enabling Big Data and online/streaming appliances with manageable time and memory footprints (FLEXA; GOMES; MOREIRA; ALVES, et al., 2021).

Taking inspiration on PCS, similar interspace-based approaches have been devised, such as the Pyramidal Embedding System (PES) of Barreto et al. (2020), which employs a polyhedron that resembles a pyramid instead of a regular polygon as its interspace.

2.3.1.4 Kernel Principal Component Analysis

Kernel-based methods (also termed kernel trick) are intensely applied in the non-linear DR literature (CHENG; LI; OGUNBONA, 2009; GOPI; PALANISAMY, 2015; KANG; PENG; CHENG, 2017; AYESHA; HANIF; TALIB, 2020). In this context, Kernel PCA (KPCA) extends PCA by means of integral operator kernel functions known from SVMs (SCHÖLKOPF; SMOLA; MÜLLER, 1997). Instead of calculating the matrix covariance, KPCA computes the principal eigenvectors of the kernel matrix, consequently extracting non-linear principal components (AYESHA; HANIF; TALIB, 2020).

Recent applications of KPCA include functional magnetic resonance imaging (TSATSISHVILI et al., 2018), nuclear accident source term estimation (LING et al., 2020), geographical discrimination of propolis (PILARIO; TIELEMANS; MOJICA, 2022), and fault monitoring of chemical processes (HAN et al., 2022).

2.3.1.5 Uniform Manifold Approximation and Projection

Introduced by McInnes, Healy, and Melville (2018) as an application of a "theoretical framework based in Riemannian geometry and algebraic topology", Uniform Manifold Approximation and Projection (UMAP) is regarded as a scalable algorithm that is suitable to perform DR while retaining global data structures. It is a non-linear, iterative routine that can perform supervised, semi-supervised and unsupervised learning.

UMAP is usually compared to t-Distributed Stochastic Neighbor Embedding (t-SNE) (MAATEN; HINTON, 2008; MAATEN, 2014) for they share a similar routine; in fact, there is even an appendix in the work of McInnes, Healy, and Melville (2018) dedicated to this discussion. However, there are distinctions between these algorithms, some of which resulted in the consideration of UMAP and the disregard of t-SNE for this

work. To begin with, UMAP begins its routine by employing a kind of Spectral Clustering method termed Laplacian Eigenmaps (BELKIN; NIYOGI, 2002, 2003; MCINNES; HEALY; MELVILLE, 2018), whereas t-SNE initializes its optimization procedure by randomly positioning the data in the projected space (VAN DER MAATEN; HINTON, 2008).

For context, the impact of different initialization techniques on UMAP and t-SNE is discussed to speed runtimes and to improve the quality of the resulting projection (KOBAK; LINDERMAN, 2019, 2021). A popular alternative initialization strategy for t-SNE is to use PCA, as done by Maaten and Hinton (2008). A variation of t-SNE termed t-Distributed Deterministic Neighbor Embedding (t-DNE), which was recently introduced by Flexa, Gomes, Moreira, Alves, et al. (2021), distinguishes itself by employing PCS to initialize the gradient descent optimization.

Another substantial distinction between both algorithms lies in their optimization procedure. UMAP makes use of cross-entropy as its cost function (MCINNES; HEALY; MELVILLE, 2018); in contrast, t-SNE resorts to the Kullback-Leibler divergence (VAN DER MAATEN; HINTON, 2008). Lastly, unlike t-SNE, UMAP can perform supervised learning and learn a mapping between the original data and the low-dimensional space, allowing the projection of unforeseen instances without re-processing the entire data set. The capacity of UMAP to produce a model was determining in including it in this work and disregarding t-SNE. These, along with other discrepancies between these algorithms beyond the scope of this work, provoke sensible differences in terms of their produced projections and runtimes.

In addition to single-cell transcriptomic studies (BECHT et al., 2019; YANG, Y. et al., 2021), UMAP applications can be found in the realms of annotation transfer between molecular imaging modalities (RACE et al., 2021), spectral artwork imaging (VERMEULEN et al., 2021), and aquatic ecology (MILOŠEVIĆ et al., 2022).

2.3.1.6 Ivis

Introduced under the context of single-cell data, Ivis is a deep-learning-based projector targeted at being an scalable routine capable of preserving both local and global features of high-dimensional data (SZUBERT et al., 2019). It is a non-linear DR technique that can be trained in a supervised, semi-supervised, or unsupervised manner. Besides single-cell data (SZUBERT et al., 2019; ISLAM; XING, 2021), there are applications of Ivis involving molecular dynamics simulations (TIAN; TAO, 2020), mass spectrometry imaging (RACE et al., 2021), chest radiography (DROZDOV et al., 2020), and other medical information (ISLAM; XING, 2021).

Its idiosyncrasies are attributed to the underlying Siamese neural network architecture it adopts, which consists of three identical neural networks that rank the similarity to the input data (TIAN; TAO, 2020), and to the loss function employed during the training

process, which is a variant of the standard triplet loss function that relies on Euclidean distances (SZUBERT et al., 2019). This loss function jointly minimizes intra-cluster distances and maximizes inter-cluster distances, leading to the understanding that Ivis preserves both local and global data arrangements in the low-dimensional space (SZUBERT et al., 2019; TIAN; TAO, 2020).

2.3.2 Classifiers

In addition to mapping observations from high-dimensional to low-dimensional spaces, there are many tasks under the ML umbrella. Among them are *classification* and *regression*, which are instances of supervised learning. The main distinction between them lies in the fact that regression produces a value in a continuous space, whereas classification labels instances based on a finite set of values (RUSSELL; NORVIG, 2021, p. 670). Classifiers are the focus of this work.

Evaluating the performance of a classification model is nontrivial. Multiple measurements exist for this task, among which the following ones:

- *Accuracy*, which is the fraction of predictions correctly predicted by a classifier.
- *F1 score*, which is the harmonic mean of the precision (i.e., positive predictive value) and recall (i.e., sensitivity) of a classifier.
- *Area under the ROC curve (AUC)*, which gauges the ability of a classifier to distinguish between classes and is used as a summary of the receiver operating characteristic (ROC) curve.
- *Matthews correlation coefficient (MCC)*, or phi coefficient, which is similar to the Pearson correlation coefficient and is a measure of the quality of binary classifications.
- *Sensitivity*, or true-positive rate, which measures how often a classifier yields a positive result for instances who have the trait being tested for.
- *Specificity*, or true-negative rate, which measures how often a classifier yields a negative result for instances who do not have the trait being tested for.

All aforementioned measurements are real-valued. Their values are in [0, 1] except for MCC, whose values are in [-1, 1]. The higher they are, the better. Differing a bit from the other measures, MCC requires a particular interpretation. The lowest MCC value indicates total disagreement between predictions and observations, whereas zero hints at a model that is no better than a random one, and the highest measurement characterizes total agreement between predictions and observations.

Table 2 – Selected supervised classification techniques ordered by ascending debut date. Source: the author.

Property / Classifier	Family[1]	Debut paper
KNN	nearest neighbor	Fix and Hodges (1951)
MLP	neural network	Werbos (1974)
SVM	support vector machine	Boser, Guyon, and Vapnik (1992)
RF	random forest	Breiman (2001)

[1] See Fernandez-Delgado et al. (2014) for more details on classifier families.

There is a wide variety of classifier families with their respective assumptions, routines and variations. Based on this and similarly to what was made for the projectors, a representative selection of classifiers was composed aiming at algorithms from different families that have competent predictive capabilities (FERNANDEZ-DELGADO et al., 2014). These classifiers, all of which seen in other articles that aim at ML-based HIA prediction (WANG, N.-N. et al., 2017; KUMAR et al., 2017; ESAKI et al., 2019) and capable of handling non-linear data, are KNN (Section 2.3.2.1), MLP (Section 2.3.2.2), RF (Section 2.3.2.4), and SVM (Section 2.3.2.3). Table 2 lists them in debut order.

2.3.2.1 *K-Nearest Neighbors*

K-Nearest Neighbors (KNN) was first defined by Fix and Hodges (1951) as a means of tackling the discrimination problem when no information on class distributions are available. As such, KNN is a nonparametric algorithm that intends to find reasonable discrimination procedures that work even if no parametric form can be assumed.

To label a given instance with the KNN procedure, as the name implies, one just needs to look at the closest instances and look for the most common label (RUSSELL; NORVIG, 2021, p. 705). This procedure assumes a definition of distance, which, although customarily defined by the Euclidean definition, can also assume other definitions (e.g., Hamming, Manhattan, Minkowski).

2.3.2.2 *Multi-Layer Perceptron*

The concept of back-propagation networks, or Multi-Layer Perceptron (MLP), was introduced by Werbos (1974) (POPOVIC, 2000). Back-propagation networks are a class of feed-forward Artificial Neural Network (ANN), which is a bioinspired ML algorithm capable of learning complex information by optimizing artificial neuron parameters called synaptic weight and bias (HAYKIN, 2008). As the name suggests, a feed-forward ANN has connections only in one direction, i.e., it can be defined as a directed acyclic graph with designated input and output nodes (RUSSELL; NORVIG, 2021, p. 802).

A MLP consists of at least three layers of nodes: an input layer, a hidden layer and an output layer. Except for the input nodes, each node is a neuron that uses a nonlinear activation function.

2.3.2.3 Support Vector Machine

Support Vector Machine (SVM) is a ML algorithm based on the statistical-learning theory (JAMES et al., 2013). It is an iteration of the so-called kernel machines. This algorithm was defined by Boser, Guyon, and Vapnik (1992) as "a training algorithm that maximizes the margin between the training patterns and the decision boundary."

SVM is a popular classification algorithm by virtue of some of its properties. Firstly, it constructs a maximum margin separator, i.e., a decision boundary with the largest possible distance to example points, thus helping in the generalization of the produced model. Furthermore, even though the hyperplanes it creates are linear, it can also classify non-linear data by resorting to the kernel trick. Lastly, the characteristics of SVM lead to its flexibility in representing complex functions while remaining resistant to overfitting (RUSSELL; NORVIG, 2021, p. 710).

2.3.2.4 Random Forest

Breiman (2001), who formalized Random Forests (RFs) as they are presently known, defined them as "a combination of tree predictors such that each tree depends on the values of a random vector sampled independently and with the same distribution for all trees in the forest." RF is an ensemble technique that utilizes several independent Decision Trees (DTs) (JAMES et al., 2013). Succinctly speaking, RF is a means of DT bagging that promotes variation between DTs, ultimately mitigating model variance.

The key idea—which doubles as the reasoning behind the name of this algorithm—is to vary the attribute choices randomly. At each split, a random sample of features is chosen, and the one that gives the highest information gain is chosen (RUSSELL; NORVIG, 2021, p. 715). RF is regarded as one of the most powerful ML algorithms (FERNANDEZ-DELGADO et al., 2014).

2.4 Hyper-parameter optimization with cross-validation

ML algorithms are seldom parameter-free (SNOEK; LAROCHELLE; ADAMS, 2012). They can have two kinds of parameter. The first type is the *model parameter*—which can be initialized and updated during the learning process—whereas the second one is the *hyper-parameter (HP)*—which cannot be estimated by the algorithm and must be set beforehand (YANG; SHAMI, 2020).

Hyper-parameter optimization (HPO) is an artisanal process, as it hinges on the employed ML algorithms and on whether a given HP is categorical, continuous, or discrete. These complications notwithstanding, executing HPO is fundamental:

> 1. It reduces the human effort required, since many ML developers spend considerable time tuning the hyper-parameters, especially for large datasets or complex ML algorithms with a large number of hyper-parameters.
>
> 2. It improves the performance of ML models. Many ML hyper-parameters have different optimums to achieve best performance in different datasets or problems.
>
> 3. It makes the models and research more reproducible. Only when the same level of hyper-parameter tuning process is implemented can different ML algorithms be compared fairly; hence, using a same HPO method on different ML algorithms also helps to determine the most suitable ML model for a specific problem (YANG; SHAMI, 2020).

Manual testing is the traditional means of performing HPO. However, in addition to requiring an expertise on the involved algorithms, this methodology does not scale well in contexts that involve many HPs, complex models, time-consuming evaluations, and non-linear HP interactions (YANG; SHAMI, 2020).

An evolution to this scenario would be to adopt a decision-theoretic procedure such as grid search (GS) (BERGSTRA et al., 2011): the specialist defines a set of possible values for each HP and the routine exhausts all possible combinations, returning the one that best optimizes the defined objective function (YANG; SHAMI, 2020). In spite of automatizing the testing process, GS still relies on the specialist to define values for each HP. Furthermore, the HPs are treated independently, meaning that all combinations are tested regardless of how promising (or not) they are. More promising strategies can be contemplated, among which the consideration of HPO through the framework of Bayesian optimization (BO) (SNOEK; LAROCHELLE; ADAMS, 2012).

BO is an iterative process that further automatizes and abstracts HPO. It is regarded as a state-of-the-art procedure for automatized ML systems (WARING; LINDVALL; UMETON, 2020). Firstly, it builds a probabilistic surrogate model of the objective function (e.g., accuracy scores), based on which the optimal HP values are detected. These values are then passed to the real objective function for evaluation, and the output of this step is used to refine the surrogate model. This process is repeated for a determined number of iterations (YANG; SHAMI, 2020). By taking advantage of a surrogate model and acting on the scores of previously-tested values, BO is not only smarter, but substantially more efficient in computational terms against GS.

BO has been employed in a variety of contexts, such as computational fluid dynamics problems (MORITA et al., 2022), molecular design and discovery (WANG; DOWLING, 2022), and optimization of ANNs to detect COVID-19 patients using computed tomography

and X-ray image data (ASLAN et al., 2022; LOEY; EL-SAPPAGH; MIRJALILI, 2022). Given its versatility, it is a pertinent choice to perform HPO of ML pipelines.

Optimization processes such as HPO are customarily guided by an objective function, also referred to as score. Accuracy is the most adopted score, although others (e.g., F1 score) can also be used. However, doing so might result in a situation where the trained model performs well on the data it saw, but not on unseen data:

> When we use data to determine some procedure, we want to be able to answer the question: How well may I expect the chosen procedure to behave in use? Even when the specific variables to be used in a multiple regression have been picked on advance, so that the form is determined, the coefficients are chosen from infinitely many combinations of possibilities to make the results of substituting in the formula fit the data as closely as possible. Testing the procedure on the data that gave it birth is almost certain to overestimate performance, for the optimizing process that chose it from among many possible procedures will have made the greatest use possible of any and all idiosyncrasies of those particular data. Sometimes we say that "Optimization capitalizes on chance!". As a result, the procedure will likely work better for these data than for almost any other data that will arise in practice. The apparent degree of fit will almost never be representative.
>
> No one knows how to appraise a procedure safely except by using different bodies of data from those that determined it. In other words, appraisal requires some form of cross-validation (JONES, 1987, p. 638).

HPO procedures can be coupled with cross-validation (CV) to rely on more representative performance measurements to guide this process. CV uses the examples of a given set for both training and validation, albeit not at the same time. This is interesting in problems whose information is scarce, as it allows for a fuller use of available information. A popular method is k-fold CV, which consists of building models on different subsets of the data such that each instance can compose either of the training or validation sets in different models. This method encompasses k rounds of learning: the training data is split into k sets, and on each round $\frac{1}{k}$ of the data are held out for validation while the remaining part is used as the training set. Ultimately, instead of a single measurement, k scores are obtained. Popular values of k are 5 and 10, as they are enough to give an estimate that is statistically likely. Note that even with CV, a separate test set is still needed (RUSSELL; NORVIG, 2021, p. 684).

2.5 Hypothesis testing for statistical significance inference

Hypothesis testing is a popular form of statistical inference that consists of inferring from a sample whether or not a given statement about the population appears to be true (CONOVER, 1999, p. 95). Following this methodology, statistical significance is attested when, given all possible outcomes (i.e., the null distribution), an obtained result is very

unlikely to have occurred. Quoting Cooper (2019), "if the outcome of an experiment deviates significantly from the most likely outcomes predicted by the null distribution, then the result is statistically significant."

The *modus operandi* of hypothesis testing involves identifying a null hypothesis, collecting data, then estimating the probability (i.e., the p-value) of getting the observed data if the null hypothesis is true. If the calculated p-value is low given the null distribution, a conclusion can be drawn that the null hypothesis probably is not true (DANIEL, 1990). How the null distribution is calculated depends on the experiment design and the collected data. Chi-square and normal distributions are commonly assumed in statistical significance experiments (COOPER, 2019).

Resorting to this kind of testing requires a conscientious methodology: the correct meaning and interpretation of its results is elusive and misinterpretations are both common and persistent (GRIFFITHS; NEEDLEMAN, 2019). Due to this, a discussion has emerged on the pertinence of its application as customarily observed in the literature (AMRHEIN; GREENLAND; MCSHANE, 2019; LOVELL, 2020). Nevertheless, statistical significance testing is frequently applied in many areas of knowledge, such as meta-heuristics (SANTOS et al., 2019, 2020; VASCONCELOS et al., 2021), ML (DEMŠAR, 2006; FERNÁNDEZ et al., 2015), and DR (FLEXA; GOMES; MOREIRA; ALVES, et al., 2021).

Depending on whether the population distribution function is known beforehand, one of two types of tests can be applied. If the distribution function is known, a parametric method can be applied. Otherwise, non-parametric methods are preferred, as they are do not assume a particular distribution. This makes this type of method valid for data from any population with any probability distribution (CONOVER, 1999, p. 115-116).

Friedman's test is one of the best-known methods in statistical analysis (AL-SWAITTI; ALBUGHDADI; ISA, 2018; GARCÍA et al., 2010). It is a non-parametric analogue of the two-way analysis of variance, and can be used to determine whether there are statistically significant differences among different techniques (DEMŠAR, 2006; PEREIRA; AFONSO; MEDEIROS, 2015). Under the null hypothesis, this test assumes that the repeated measures or matched groups come from the same population or from populations with the same median (PEREIRA; AFONSO; MEDEIROS, 2015).

If the null hypothesis is rejected and the existence of significant differences between treatments is assumed to be true, a *post-hoc* procedure (e.g., Conover, Nemenyi) can be conducted to pinpoint which treatments differ from others (PEREIRA; AFONSO; MEDEIROS, 2015). This comprises comparing all possible method pairs, something that incurs the rejection of a certain proportion of the null hypothesis by chance. This phenomenon is called family-wise error (FWE), a well-known statistical problem for which different adjustment techniques are available (e.g., Bonferroni, Hommel) (DEMŠAR, 2006).

An important note when evaluating p-values is that absence of evidence does not imply evidence of absence, i.e., no evidence of effect is not equal to evidence of no effect:

> A high P-value does not disprove the possibility that there is a true difference in the population, but merely indicates that our sample data do not demonstrate one with certainty. In other words, we cannot equate a failure to reject the null hypothesis with acceptance of the null hypothesis (ROBINSON; HAVILAND, 2021).

Likewise, a low p-value does not prove, but rather indicates how strongly the data imply that a difference exists. p-values can never be 0 or 1; any such reports indicate that the p-values were rounded or truncated (ROBINSON; HAVILAND, 2021).

CHAPTER

3

MATERIALS AND METHODS

Following the contextualization and definition of relevant concepts promoted by previous chapters, this one describes the data set produced for this work, the methodology behind the production of the pipelines whose results are scrutinized hereafter, and the computational devices, programming languages, and tools used in this process.

It is important to accentuate the existence, other than the data set production procedure, of two distinct processes herein. The pipelines *per se* execute a process comprising multiple estimators to project and classify an example; however, another procedure is carried out beforehand to build and optimize them. This earlier process, which is also referred to henceforth as the *meta-process*, is also used for the purposes of this work to extract performance measurements and compare the produced and optimized pipelines.

All processing is performed in the `Python` programming language, version `3.9.5`, running on a Dell PowerEdge T440 server powered by an Intel® Xeon® Silver 4114 CPU @ 2.20 GHz with ten cores and twenty threads, as well as 32 GB of DDR4-2400 RAM. Canonical® Ubuntu® 20.04.3 LTS (Focal Fossa) is used as the operating system. All plots are made with the `matplotlib` (HUNTER, 2007), `pandas` (MCKINNEY et al., 2010), and `seaborn` (WASKOM, 2021) modules, versions `3.4.3`, `1.3.4`, and `0.11.2` respectively.

Figure 2 describes the aforementioned processes while highlighting some of the modules used in them. The first two blocks (i.e., the ones in blue and green, from left to right) respectively describe the processes that generate the data set used as input and the results outputted by the pipelines. In specific terms, the data set generation process in blue effectively generates the data set to be used from the small molecules to be analyzed. This data set is used as input for the pipeline generation process illustrated in the green block, which produces trained pipelines with optimized HPs. Once fitted, these pipelines can perform predictive tasks, viz., projecting and classifying unseen small molecule data.

As illustrated in the brown and orange blocks at the right-hand side of Figure 2,

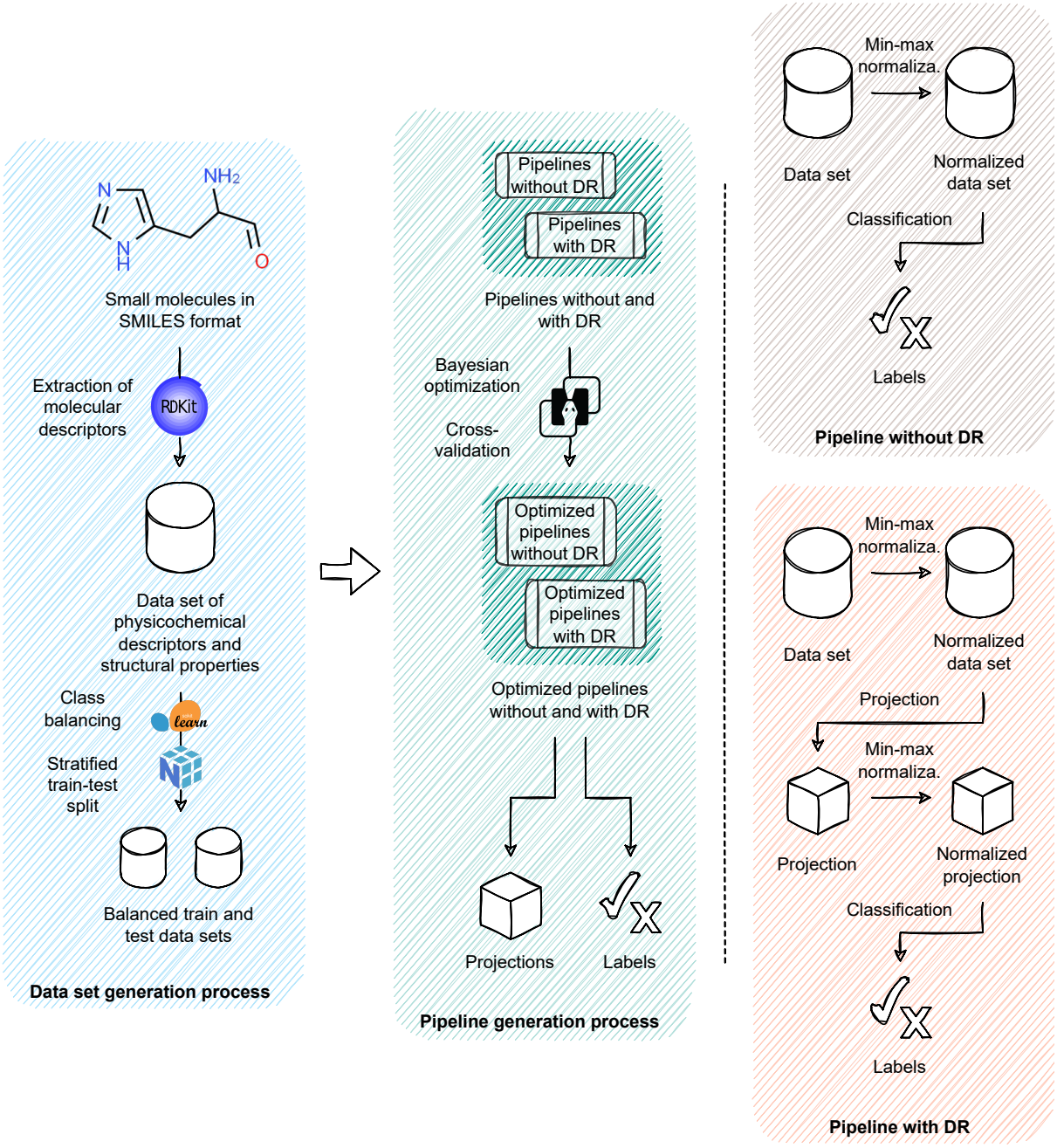


Figure 2 – Steps of the data set and pipeline generation processes, as well as those of pipelines without and with DR. Source: the author.

pipelines without and with DR are produced herein. This allows one to gauge the impact of projecting data on the prediction performance, ultimately enabling an inspection on which pipelines best fit the train data assuming that the same input information is used.

The sections of this chapter are dedicated to clarifying the processes and pipeline structures illustrated in Figure 2. The data set generation is described in Section 3.1, whereas the pipeline production meta-process and the processes for pipelines without and with DR are respectively detailed in Sections 3.2 and 3.3.

3.1 Data set generation process

The data set generation procedure begins with a representation of relevant small molecules. For this work, an unbalanced data set is obtained from Ning-Ning Wang et al. (2017). The compounds that compose this set are described as strings that follow a chemical notation named Simplified Molecular Input Line Entry System (SMILES). It contains 152 compounds with acceptable permeability and 818 poorly-absorbed ones. They will be regarded henceforth as HIA (+) and HIA (–) molecules. An undersampling of the dominant class is carried out with the `scikit-learn` module, version `1.0.1` (PEDREGOSA et al., 2011), resulting in two balanced classes comprising 152 chemicals each. This is performed by randomly excluding examples of the dominant class until balance is reached.

Afterwards, numerical features are generated from the attained SMILES strings. This procedure, which is performed with the `RDKit` module, version `2020.09.1.0` (LANDRUM, 2020), results in a collection of fifty physicochemical descriptors and structural properties. No fingerprints are used. The selected descriptors include the previously-mentioned $Fcsp^3$, HBA, HBD, logP, MW, NRB, and tPSA. The remaining features are:

- number of aromatic rings;
- aliphatic carboxylic acids;
- aliphatic hydroxyl group;
- N functional groups attached to aromatics;
- aromatic carboxylic acids;
- aromatic nitrogens;
- aromatic amines;
- aromatic hydroxyl;
- carboxylic acids;
- carbonyl O;
- thiocarbonyl;
- imines;
- primary amines;
- secondary amines;
- tertiary amines;
- hydroxylamine;
- thiol;
- aldehydes;
- amides;
- amidine;
- anilines;
- azide groups;
- azo groups;
- benzene rings;
- bicyclic;
- esters;
- guanidine;
- halogens;
- ketones;
- nitriles;
- nitro groups;
- phenols;
- urea groups;
- primary amides;
- ether oxygens;
- furan rings;
- thioether;
- sulfonamides;
- sulfone groups;
- quaternary nitrogen;
- methoxy groups;
- beta lactams; and
- cyclic esters (lactones).

Lastly, a stratified train-test-split is performed with `scikit-learn`, resulting in a train and test set with 212 and 92 instances respectively, both with balanced class sizes. Further manipulation is performed with the `numpy` module, version `1.19.5` (HARRIS et al., 2020), and `pandas`, to produce data sets that are suitable for subsequent consumption.

3.2 Pipeline production process

This meta-process is where all combinations of projectors and classifiers are separately created and fine-tuned. Two tasks are performed. The first, denominated pipeline tuning, involves the creation of multiple instances of a given pipeline with different HPs in a search procedure. This HPO procedure culminates with the attainment of the best pipelines, i.e., the HP settings among those evaluated that best optimize the objective function. In addition to performance measurements, runtimes are also registered to evaluate how long it takes to fit and predict each pipeline. Statistical significance and post-hoc testing is performed to determine how likely a given discrepancy is to occur assuming that any observed differences between pipelines are irrelevant.

The second phase, termed independent testing, further scrutinizes the fifty-two best-performing pipelines found in the previous step by gauging their predictive capabilities on unseen data. The produced projections are visually inspected and multiple measures are obtained from the labels predicted for the hold-out set, allowing a holistic analysis and comparison of the pipelines in terms of exploratory data analysis and HIA prediction.

Details on the pipeline tuning and independent testing phases are respectively located in Sections 3.2.1 and 3.2.2.

3.2.1 Pipeline tuning

This phase, which only uses the train set, is performed via BO. The fifty-two pipelines considered herein are individually tweaked in a process performed with the `BayesSearchCV` estimator of the `scikit-optimize` module, version `0.9.0`, aiming to maximize classification accuracy on unseen data. This HPO procedure is coupled with 10-fold CV to also use the train set for validation purposes, meaning that the accuracy that guides the optimization procedure is obtained by averaging ten executions over different train-validation splits of the train set.

For each pipeline, BO is performed for one hundred iterations that evaluate ten HP settings each. Considering that each HP evaluation comprises ten models as part of the adopted CV method for fifty-two pipelines, fifty-two thousand HP settings are evaluated and 520,000 models are generated globally via a script whose execution customarily lasts around two days. In spite of the sequential nature of BO (YANG; SHAMI, 2020), the

evaluation of the points within iterations is parallelized in this work by means of the `n_jobs` HP to take advantage of the available computational resources.

It is important to highlight that, since the same number of evaluations is performed for all pipelines regardless of the number of HPs to optimize, pipelines with a lesser number of HPs tend to be favored. The reasoning behind this is that pipelines with a bigger number of HPs have to optimize them with the same number of iterations used to optimize pipelines with a lower-dimensional HP space.

Table 3 describes the search space of the HPO process for the considered projectors and classifiers. The HP search space of pipelines with DR is the junction of the ones of their comprising projector and classifier. HPs that are categorical and only have one value signal that their default setting was changed to another one, which stays constant throughout the optimization procedure. The remaining HPs of these estimators, if existent, remained in their respective default values. They can be consulted by referring to the documentation of the employed implementations.

In terms of used implementations, all classifiers, as well as the PCA, TSVD, and KPCA projectors, are from `scikit-learn`. The UMAP implementation is from the `umap-learn` module, version `0.5.2` (MCINNES; HEALY; MELVILLE, 2018), whereas the Ivis one is from the `ivis` module[1] (SZUBERT et al., 2019). PCS was adapted from an implementation disclosed by Barreto et al. (2020).

Whenever possible, the estimators are set to parallelize their execution via the `n_jobs` HP, thus taking advantage of all available threads. If available, the estimators with stochastic routines are provided with a seed to their random number generators by means of the `random_state` HP to make the experiment more reproducible and the HP evaluations more comparable. This seeding procedure is unavailable in Ivis due to implementation limitations, meaning that its results might slightly vary upon retraining even if the same data set and HPs are used.

Some particularities about the chosen HP search space are worth mentioning. For MLP, due to implementation inconveniences, the number of hidden layers is fixed at two and only their comprising number of neurons is tuned. The `max_iter` HP of SVM is set to a reasonably high value to establish a roof and interrupt the execution of HP settings that take unreasonably long runtimes to converge. Additionally, for UMAP and Ivis—which are capable of supervised learning—the weight of the labels is tuned by means of the `target_weight` and `supervision_weight` HPs, respectively. This shall allow the inspection of how important label information ultimately is according to the search of the

[1] This work required functionality that was not present in version `2.06` of `ivis`, the latest one at the time this manuscript was written. The specific implementation can be installed from the @beringresearch/ivis repository at GitHub by running this command: `pip install git+git://github.com/beringresearch/ivis.git@57bfe6f0673b0b0e6cc8c30a4b6b3772e3ab9b08`.

Table 3 – Search space of the considered hyper-parameters for all estimators of the projection and classification phases. Source: the author.

Search space		Hyper-parameter	Type	Interval
Phase	Estimator			
projection	PCA	—	—	—
	TSVD	`algorithm`	categorical	[`arpack`]
	PCS	—	—	—
	KPCA	`eigen_solver`	categorical	[`dense`]
		`gamma`	real	[0, 1]
		`kernel`	categorical	[`cosine`, `linear`, `rbf`, `sigmoid`]
		`n_jobs`	categorical	[`-1`]
	UMAP	`n_epochs`	integer	[100, 2000]
		`n_neighbors`	integer	[100, 2000]
		`random_state`	categorical	[`2021`]
		`target_weight`	real	[0, 1]
	Ivis	`k`	integer	[3, 103]
		`model`	categorical	[`hinton`, `maaten`, `szubert`]
		`n_epochs_without_progress`	integer	[10, 50]
		`supervision_weight`	real	[0, 1]
classification	KNN	`leaf_size`	integer	[3, 30]
		`n_neighbors`	integer	[3, 30]
		`p`	integer	[1, 5]
		`weights`	categorical	[`distance`, `uniform`]
	MLP	`activation`	categorical	[`logistic`, `relu`, `tanh`]
		`alpha`	real	[1e-4, 1e-2]
		`hidden_layer_sizes`	(integer, integer)	([50, 150], [50, 150])
		`max_iter`	integer	[100, 2000]
		`random_state`	categorical	[`2021`]
	SVM	`C`	real	[1e-3, 1e3]
		`gamma`	real	[1e-2, 1e2]
		`kernel`	categorical	[`linear`, `rbf`, `sigmoid`]
		`max_iter`	categorical	[`100000`]
		`probability`	categorical	[`True`]
		`random_state`	categorical	[`2021`]
	RF	`max_depth`	integer	[3, 15]
		`n_estimators`	integer	[50, 150]
		`n_jobs`	categorical	[`-1`]
		`random_state`	categorical	[`2021`]

BO for pipelines involving these projectors. Lastly, the `gamma` HP that is present in the KPCA projector and the SVM classifier only affects certain kernels and is ignored when the `cosine` or `linear` settings are in place for the `kernel` HP.

The mean train and validation scores of the best pipelines are obtained as products of the pipeline tuning procedure, as well as the mean fit and predict runtimes. The mean train score is used to see how well the pipelines could predict previously-seen molecules, whereas the mean validation score serves to inspect how well the pipelines sustained their performance for unforeseen compounds. The runtimes can be inspected to see how long it takes for a given pipeline to be trained and to predict the labels of the validation sets. Additionally, the HPs that the best-performing models use can also be analyzed to

rationalize the directions that the HPO procedure took for each pipeline.

A holistic evaluation of the produced pipelines must comprise both performance scores and execution times. Although runtimes might seem tolerable for the employed data set and techniques, any observed differences should be assumed to accentuate as the employed data grows. Furthermore, these discrepancies compound in HPO procedures, where multiple models are evaluated, and can sensibly protract their execution. An ideal pipeline should have maximum mean accuracy and minimum mean runtime.

Lastly, the Friedman's statistical significance test is performed on the train and validation scores to ascertain, based on the results of the 10-fold CV, whether the observed performance differences are meaningful between DR-encompassing pipelines and those trained on the fifty-dimensional data set. The null hypothesis under scrutiny is that there is no statistically significant accuracy discrepancy between pipelines. The `scipy` module, version `1.7.1` (VIRTANEN et al., 2020), is used to perform these tests.

Whenever statistical significance is attested in Friedman's test results, Conover and Nemenyi post-hoc tests are executed with the `scikit-posthocs` module, version `0.6.7` (TERPILOWSKI, 2019). Since the p-values yielded by the Conover procedure, unlike those of Nemenyi, are not adjusted to account for the FWE, the Bonferroni and Hommel adjustment methods are applied as well. The joint inspection of these pairwise tests shall pinpoint the pipeline pairs whose performance differed substantially, therefore allowing a reflection on the reasons behind this.

3.2.2 Independent testing

After finding the best HPs for each pipeline and studying their performance in terms of mean accuracy and runtime, this phase makes use of the test set to perform a last experiment. The goal is to see how successful the previous phase was in obtaining pipelines that modeled the problem while disregarding idiosyncrasies of the train set.

There are two main products of this phase. The first one is the projections predicted by the pipelines on the test set. Some projectors have neither stochastic routines nor HPs to tune, meaning that their projections should be the same across pipelines. Contrastingly, others might yield different projections across pipelines, as they are being optimized to maximize the performance of the classifier alongside which they are. As such, visually inspecting the produced projections for all pipelines might give some insights on the results of the performed HPO. Furthermore, jointly inspecting different projections of a given set allows a holistic view of the data: if techniques with different intuitions and routines depict similar structural patterns for HIA (+) and HIA (–) small molecules, this might indicate that these patterns are present in the original feature space as well.

The second product is the measurements of the classification performance on the

test set for pipelines without and with DR. In this analysis, six scores are obtained: accuracy, F1, AUC, MCC, sensitivity, and specificity. Based on these measurements, a global analysis can be made on the predictive capability of the DR-encompassing pipelines against their baseline counterparts to see if applying DR via feature extraction is applicable in predicting HIA of small molecules.

3.3 Pipeline process

As a product of the pipeline production process detailed in the previous section, optimized pipelines without and with DR are created. Table 4 exposes, in order of use, the estimators that compose each pipeline type. In short, one does not have the projection step, whereas the other has. Since pipelines devoid of DR directly use the data set to train classifiers, only one normalization is needed in them.

Table 4 – Breakdown of the estimators of each step for pipelines without and with DR. Source: the author.

Pipeline / Estimator	Without DR	With DR
1	Normalizer	Normalizer
2	Classifier	Projector
3	—	Normalizer
4	—	Classifier

The pipelines are made such that the only estimators subject to change are those responsible for projection and classification. A total of six projectors (PCA, TSVD, PCS, KPCA, UMAP, and Ivis) and four classifiers (KNN, MLP, SVM, and RF) are considered in this work. Considering that all projector-classifier combinations are evaluated, there are four baseline pipelines (i.e., four pipelines without DR) and twenty-four DR-encompassing pipelines. This work covers two- and three-dimensional projections, which doubles the number of pipelines with DR to forty-eight. Overall, fifty-two pipelines are considered.

Min-max normalization is performed (also by resorting to `scikit-learn`) before the projection and classification steps to ensure that the estimators in these steps always see the data around the $[0, 1]$ interval. Two normalization procedures are performed in DR-encompassing pipelines due to the fact that some projectors, due to the particularity of their routines, might produce projections in a substantially different scale than the one of the original set. The normalization estimators learn the upper and lower boundaries of each feature in the training phase and resort to this knowledge when normalizing unseen data. Examples with values exceeding the learned boundaries are not truncated, meaning that values outside the normalized interval are possible and this interval, although promoted, is not enforced.

CHAPTER

4

RESULTS AND DISCUSSION

Based on the materials and methods described in the previous chapter, this one exposes and discusses the obtained results for the pipeline tuning (Section 4.1) and independent testing (Section 4.2) phases.

Whenever possible, table cells were colored based on the values of its pertaining column on a gradient that begins with red, goes through yellow, and ends with green. The redder a given cell is, the worse its value is; conversely, the greener it is, the better its value is. Intermediate values relative to the value range of a given column assume yellower tones. For tables pertaining to statistical significance *post-hoc* test results, cells in red contain values below the significance level and highlight pairwise comparisons regarded as statistically significant.

4.1 Pipeline tuning

The analyses of the products of the pipeline tuning phase are divided in three parts: CV measurements of the pipelines with the best-performing HP settings (Section 4.1.1), statistical significance analysis with Friedman's test and *post-hoc* testing (Section 4.1.2), and inspection of the encountered HPs (Section 4.1.3).

4.1.1 Cross-validation measurements

Table 5 exposes the train and validation accuracy scores, as well as the fit and predict runtimes (i.e., time periods used to train and validate a model, respectively) in seconds averaged from the results of the performed 10-fold CV procedure.

A first glance allows us to see the worst performers in terms of accuracy scores. MLP classifiers trained on the original data set, as well as on two-dimensional UMAP and Ivis projections, obtained measurements around 50% already on the training phase.

Table 5 – Mean train/validation accuracy scores and fit/predict execution times of best-performing pipelines in 10-fold stratified CV. Source: the author.

		Measure	Train accuracy	Validation accuracy	Fit runtime	Predict runtime
Dimensionality	Projector	Classifier				
50	—	KNN	0.7406 ± 0.0170	0.6571 ± 0.1135	0.0051 ± 0.0016	0.0100 ± 0.0036
		MLP	0.5021 ± 0.0011	0.4810 ± 0.0100	2.0270 ± 0.1369	0.0027 ± 0.0036
		SVM	0.9177 ± 0.0126	0.7784 ± 0.0828	0.0985 ± 0.0258	0.0013 ± 0.0002
		RF	0.9990 ± 0.0022	0.8253 ± 0.0708	0.2146 ± 0.0203	0.0356 ± 0.0184
2	PCA	KNN	1.0000 ± 0.0000	0.7078 ± 0.0710	0.0080 ± 0.0017	0.0020 ± 0.0002
		MLP	0.7060 ± 0.0256	0.6931 ± 0.0793	0.3705 ± 0.0289	0.0018 ± 0.0004
		SVM	0.7982 ± 0.0175	0.6842 ± 0.0675	0.1437 ± 0.0694	0.0018 ± 0.0013
		RF	0.7867 ± 0.0182	0.7455 ± 0.0695	0.2543 ± 0.0215	0.0228 ± 0.0013
	TSVD	KNN	1.0000 ± 0.0000	0.6597 ± 0.0912	0.0092 ± 0.0012	0.0022 ± 0.0002
		MLP	0.6640 ± 0.0190	0.6563 ± 0.0759	0.4677 ± 0.0278	0.0022 ± 0.0002
		SVM	0.8249 ± 0.0309	0.6457 ± 0.0976	0.1421 ± 0.0433	0.0019 ± 0.0017
		RF	1.0000 ± 0.0000	0.6794 ± 0.0521	0.6125 ± 0.0334	0.0490 ± 0.0063
	PCS	KNN	0.6955 ± 0.0242	0.6364 ± 0.1465	0.0086 ± 0.0020	0.0070 ± 0.0017
		MLP	0.6672 ± 0.0108	0.6223 ± 0.1086	0.3889 ± 0.0633	0.0067 ± 0.0011
		SVM	0.7107 ± 0.0184	0.6504 ± 0.1245	0.0570 ± 0.0118	0.0052 ± 0.0003
		RF	1.0000 ± 0.0000	0.5755 ± 0.0762	0.2900 ± 0.0357	0.0396 ± 0.0139
	KPCA	KNN	0.7479 ± 0.0144	0.7357 ± 0.0748	0.0570 ± 0.0086	0.0279 ± 0.0039
		MLP	0.7055 ± 0.0144	0.6896 ± 0.0979	0.3445 ± 0.0376	0.0490 ± 0.0292
		SVM	0.7097 ± 0.0148	0.6991 ± 0.0975	0.1147 ± 0.0110	0.0425 ± 0.0123
		RF	0.9995 ± 0.0017	0.7080 ± 0.0703	0.2794 ± 0.0402	0.0659 ± 0.0175
	UMAP	KNN	1.0000 ± 0.0000	0.7608 ± 0.0896	2.4852 ± 0.2446	1.5056 ± 0.2197
		MLP	0.5000 ± 0.0025	0.5000 ± 0.0224	3.7267 ± 0.1962	1.8610 ± 0.1700
		SVM	0.7086 ± 0.0767	0.6567 ± 0.1243	1.9406 ± 0.2242	1.4486 ± 0.2210
		RF	1.0000 ± 0.0000	0.7742 ± 0.0629	2.6243 ± 0.2281	1.3782 ± 0.2116
	Ivis	KNN	1.0000 ± 0.0000	0.8310 ± 0.0684	263.8681 ± 76.3864	0.7927 ± 0.2518
		MLP	0.5021 ± 0.0011	0.4810 ± 0.0100	32.2996 ± 7.6062	0.5360 ± 0.1722
		SVM	0.9906 ± 0.0162	0.8260 ± 0.0909	333.0231 ± 124.4573	0.8614 ± 0.2533
		RF	1.0000 ± 0.0000	0.8210 ± 0.0762	327.53 ± 94.6619	1.4752 ± 0.1677
3	PCA	KNN	0.7684 ± 0.0159	0.7314 ± 0.0812	0.0082 ± 0.0015	0.0032 ± 0.0007
		MLP	0.7427 ± 0.0116	0.7409 ± 0.0795	1.5443 ± 0.0973	0.0021 ± 0.0001
		SVM	0.9287 ± 0.0121	0.7022 ± 0.0886	0.0928 ± 0.0220	0.0013 ± 0.0002
		RF	0.9817 ± 0.0090	0.7736 ± 0.0698	0.2534 ± 0.0228	0.0315 ± 0.0059
	TSVD	KNN	1.0000 ± 0.0000	0.7268 ± 0.0813	0.0078 ± 0.0012	0.0022 ± 0.0003
		MLP	0.7128 ± 0.0262	0.6989 ± 0.0761	0.5159 ± 0.0649	0.0023 ± 0.0009
		SVM	0.7070 ± 0.0255	0.7128 ± 0.0724	0.0176 ± 0.0033	0.0015 ± 0.0003
		RF	0.8496 ± 0.0220	0.7177 ± 0.0857	0.4203 ± 0.0352	0.0527 ± 0.0193
	PCS	KNN	1.0000 ± 0.0000	0.6409 ± 0.1108	0.0096 ± 0.0011	0.0056 ± 0.0009
		MLP	0.5000 ± 0.0025	0.5000 ± 0.0224	1.2142 ± 0.0869	0.0058 ± 0.0017
		SVM	0.5116 ± 0.0190	0.4855 ± 0.0224	0.0186 ± 0.0033	0.0055 ± 0.0002
		RF	1.0000 ± 0.0000	0.6032 ± 0.0768	0.2046 ± 0.0263	0.0322 ± 0.0075
	KPCA	KNN	1.0000 ± 0.0000	0.7366 ± 0.0892	0.0774 ± 0.0144	0.0364 ± 0.0114
		MLP	0.7516 ± 0.0192	0.7221 ± 0.1040	0.4898 ± 0.0896	0.0650 ± 0.0300
		SVM	0.7269 ± 0.0122	0.7275 ± 0.1194	0.0261 ± 0.0035	0.0328 ± 0.0145
		RF	1.0000 ± 0.0000	0.7268 ± 0.1054	0.4329 ± 0.0186	0.0539 ± 0.0036
	UMAP	KNN	1.0000 ± 0.0000	0.7697 ± 0.0598	3.5638 ± 0.2369	1.4701 ± 0.1589
		MLP	0.5000 ± 0.0025	0.5000 ± 0.0224	2.5560 ± 0.2265	2.0631 ± 0.1659
		SVM	0.6021 ± 0.2097	0.4991 ± 0.0484	5.0432 ± 0.5012	1.8808 ± 0.2254
		RF	1.0000 ± 0.0000	0.6894 ± 0.0739	4.7565 ± 0.1884	1.6826 ± 0.1002
	Ivis	KNN	0.9466 ± 0.0212	0.8219 ± 0.0753	255.3325 ± 49.0265	0.5076 ± 0.2948
		MLP	0.9906 ± 0.0095	0.8403 ± 0.0692	1.8062 ± 0.7161	0.3298 ± 0.3063
		SVM	0.9911 ± 0.0082	0.8545 ± 0.0765	1.8930 ± 0.3650	0.6887 ± 0.2262
		RF	0.9995 ± 0.0017	0.6842 ± 0.0576	91.7132 ± 15.638	0.3157 ± 0.1490

This also happened to the models trained on three-dimensional PCS and SVM projections. The low standard deviation attests that this behavior is recurrent, leading to questions behind this result. Given that each pipeline was the best-performing one in terms of HP settings, as they were chosen out of a thousand tested combinations chosen by an intelligent optimization procedure, and how this happened for a variety of projectors and the original data set, MLP seems to be the underperforming estimator. Two possibilities come to

mind. One is that the data is insufficient in terms of size or is difficult to be learned by this classifier. Another one is that the `hidden_layer_sizes` HP might need to have its search space vary in terms of both the number of hidden layers and their respective size. On fewer occasions, SVM pipelines also performed similarly bad, thus warranting further investigation on MLP and SVM. Given that the data set is balanced, one hypothesis is that MLP decided that always predicting the same class is pertinent. If this happens in the independent testing phase, sensitivity and specificity scores should help verify this.

Looking at the runtimes, pipelines comprising Ivis models seem to have particularly prolonged fit times, as well as substantially longer predict times. UMAP-encompassing pipelines take even longer to predict than Ivis overall but are substantially faster to fit. Nevertheless, using UMAP and Ivis incur relatively protracted fit and predict times.

Figures 3 and 4 portray the accuracy scores observed in Table 5, albeit stratified first by classifier instead of by projector. They facilitate direct comparisons between DR-encompassing pipelines and their baseline counterparts. In terms of train accuracy, Figure 3 leads to the understanding that, except for MLP due to its previously-observed issues, all classifiers had at least one DR-encompassing pipeline that surpassed baseline performance. KNN pipelines with DR are arguably the best-performing ones in this phase when compared to the one trained on the original train set. Contrariwise, the scenario

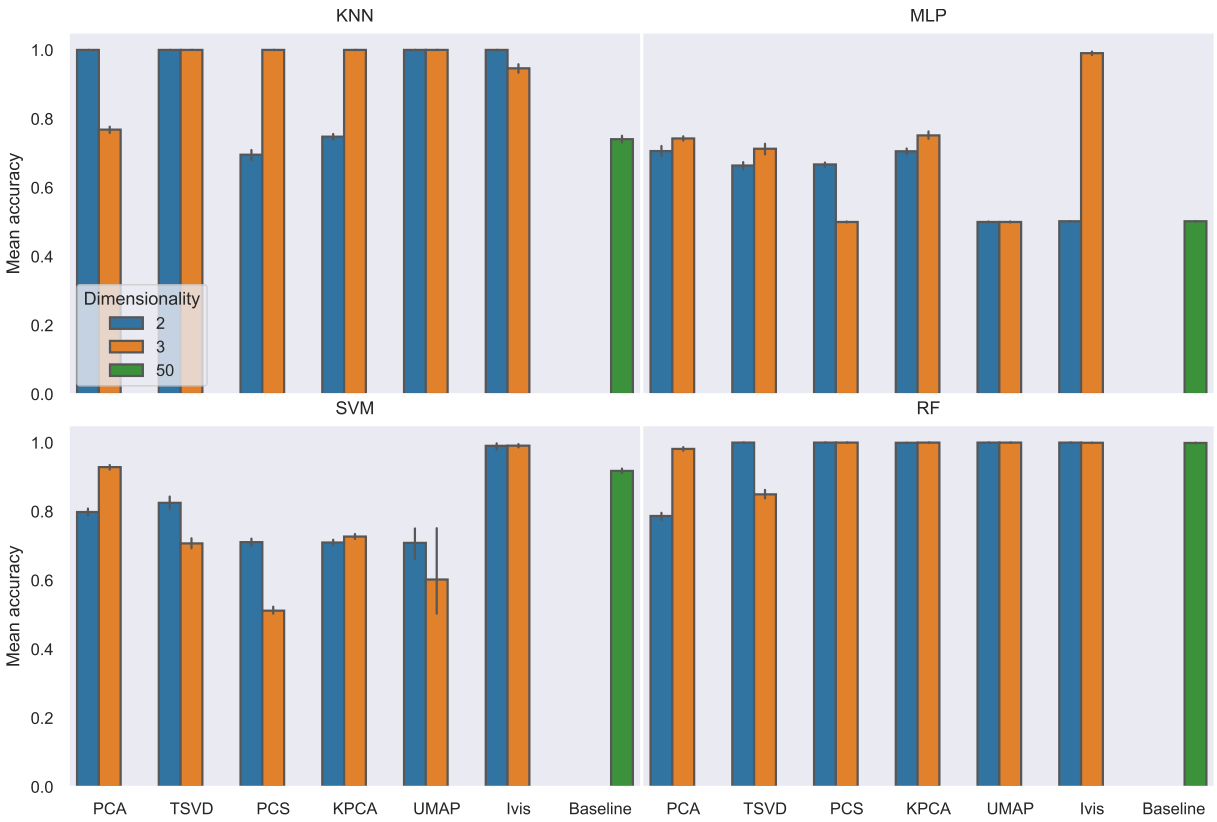


Figure 3 – Mean train accuracy of best models for each classifier grouped by projector and stratified by dimensionality. Source: the author.

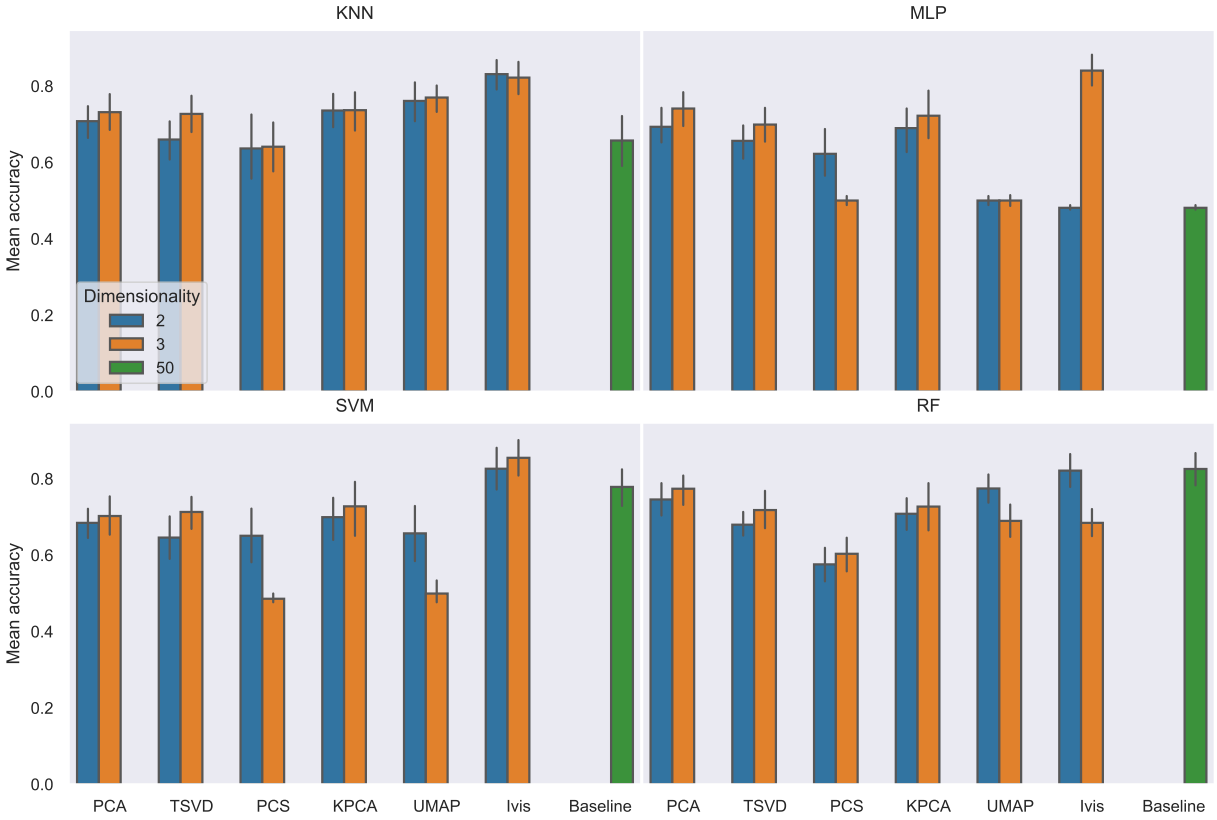


Figure 4 – Mean validation accuracy of best models for each classifier grouped by projector and stratified by dimensionality. Source: the author.

is not as favorable for MLP and SVM pipelines with DR. Lastly, both types of pipeline performed comparably for RF here. Considering what was discussed above, this certainly justifies investigating what provoked this and devising forms to overcome these hurdles.

Figure 4 is the most relevant one, as it enables inspection on how the pipelines performed on unforeseen data in the validation phase. As expected, a certain performance loss is observed across the board. Interestingly, Ivis-based pipelines sustained their performance compared to the baseline ones overall. Notably, RF pipelines with DR are not as competitive against the baseline one as they were in Figure 3, something that might indicate that, despite the employed CV procedure, the trained models still overfitted.

UMAP and Ivis tend to take the lead over the remaining projectors in terms of mean validation accuracy, something that could be justified by the possibility they have of resorting to label information. In particular, Ivis-based pipelines surpass pipelines comprising other projectors for at least one dimensionality and are competitive against the baseline ones. This notwithstanding, it is interesting to see how some techniques that can only perform unsupervised learning and apply linear transformations do not lag behind by much. The epitome of this is PCA, a classic approach that is still widely used today for its versatility, lack of HPs, and interpretability.

Figures 5 and 6, on average, confirm what was described in Section 2.3.1: non-

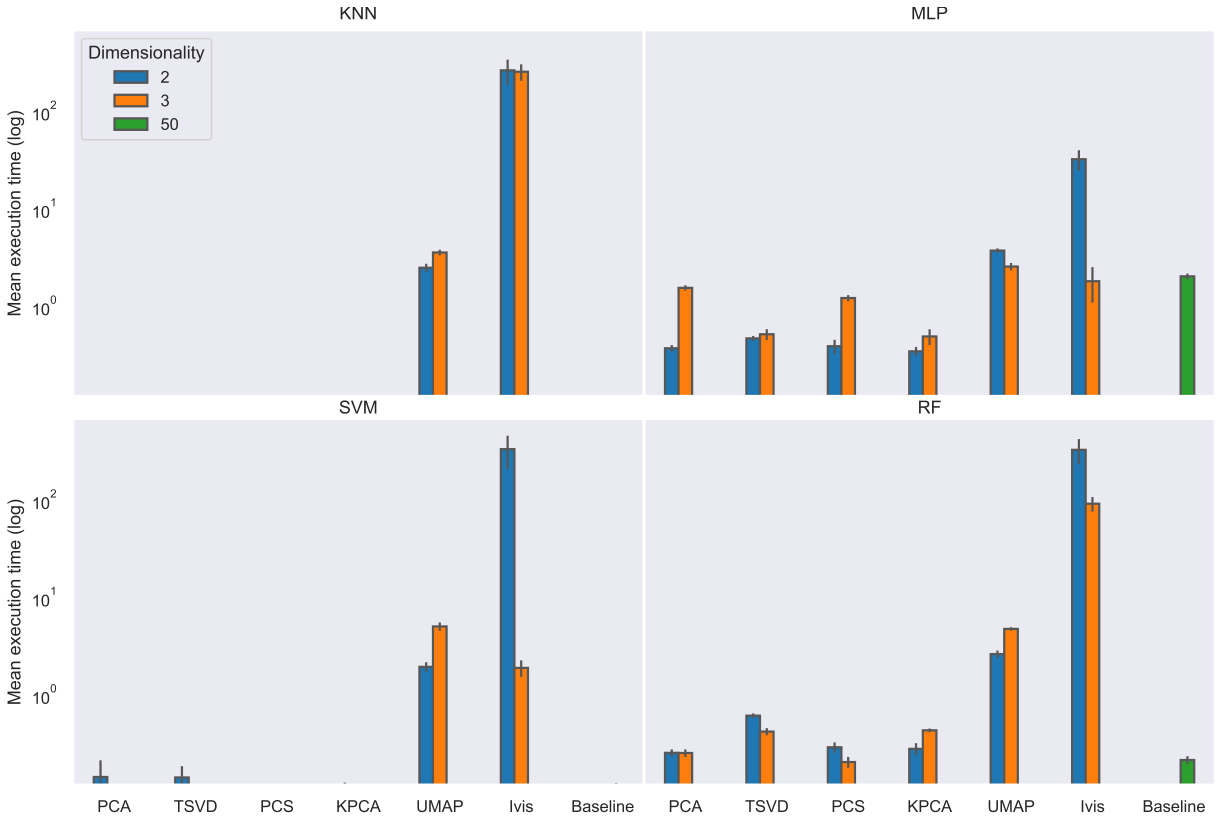


Figure 5 – Mean fit time of best models for each classifier grouped by projector and stratified by dimensionality. Source: the author.

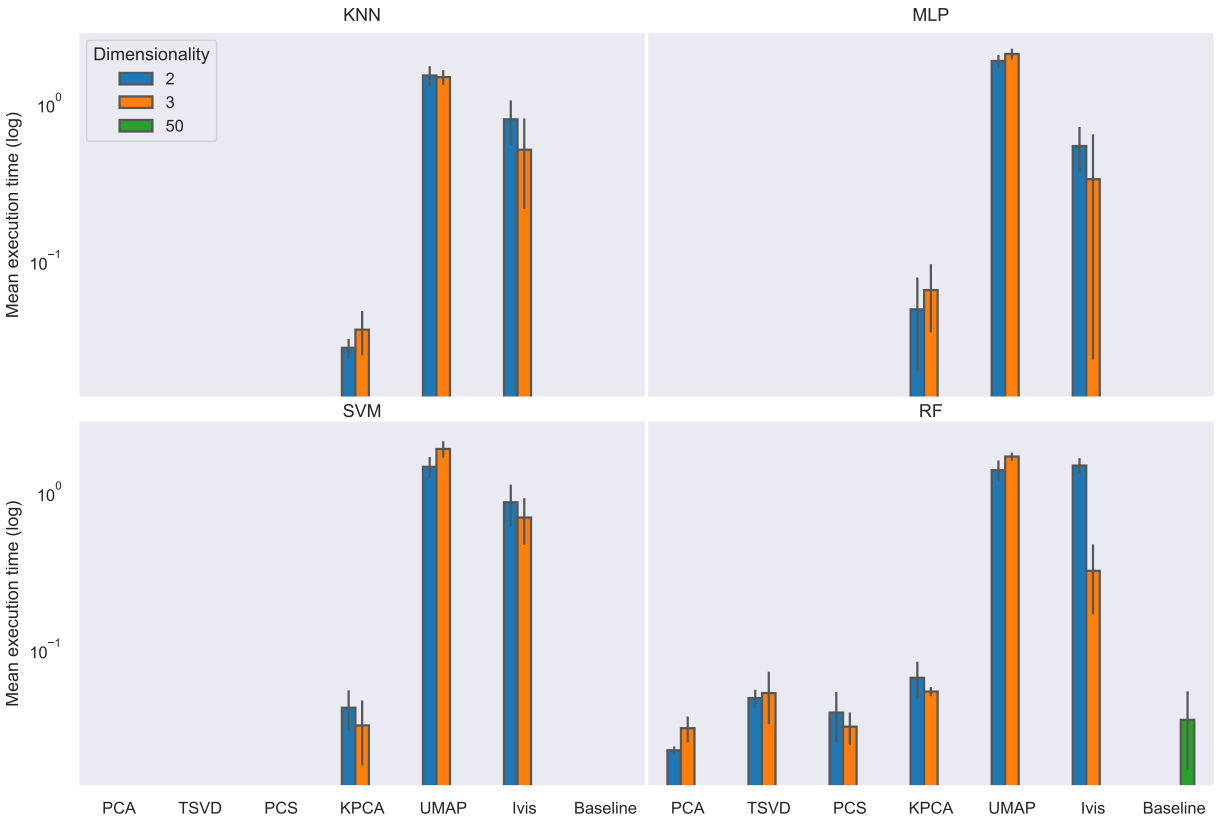


Figure 6 – Mean predict time of best models for each classifier grouped by projector and stratified by dimensionality. Source: the author.

linear projectors tend to be computationally costlier than the ones that apply linear transformations. Notice how the mean execution time in these figures is on a logarithmic scale. Interpreting them leads to the understanding that UMAP and Ivis take several orders of magnitude longer to fit than PCA, TSVD, and PCS. This is a factor worth considering, especially in other applications where sizable data sets are involved.

Another noteworthy observation is that virtually all DR-encompassing pipelines take longer to fit than pipelines devoid of DR estimators. Shorter runtimes are expected for the baseline pipelines, since the train set has little over two hundred instances and fifty dimensions. This means that, on the one hand, using pipelines with DR incurs additional overhead that also needs to be accounted for if the data set is small. On the other hand, the training times are low in absolute terms precisely because of the small size of the data, meaning that applying DR via feature extraction is feasible and might be worth it if it improves predictive performance and if visual inspection of the data for exploratory analysis is pertinent.

Complementing Table 5, measurements of best-performing pipelines obtained in each fold of the CV procedure can be found in Appendix A.

4.1.2 Statistical significance analysis

Table 6 exposes the result of Friedman's test applied on the train and validation accuracy scores, adopting a significance level of 0.05. The obtained p-values are considerably smaller than the significance level, and the Friedman statistic is many times superior to the calculated critical value. Therefore, discrepancies that are unlikely to occur by chance were detected, consequently warranting the execution of *post-hoc* testing to see for which pipeline pairs the null hypothesis is unlikely to hold true.

Table 6 – Friedman statistical significance results for train and validation accuracy scores.

Measure / Procedure	Significance level	p-value	Critical value[1]	Friedman statistic
Train	0.05	1.022026e-72	68.669294	488.357982
Validation		7.821706e-41		321.100080

[1] The critical value was obtained from a Chi-square distribution assuming a confidence level of 0.95 (i.e., one minus the significance level) and fifty-one degrees of freedom (i.e., the number of compared pipelines minus one).

For the *post-hoc* tests, all possible pipeline pairs were compared. In this section, Tables 7 and 8 contain the p-values of comparisons between pipelines encompassing the same classifier without and with DR in the training and validation phases respectively, allowing a direct inspection on how discrepant the performance of DR-encompassing pipelines is against their baseline counterparts assuming that the null hypothesis is true.

Table 7 – Friedman *post-hoc* tests on train accuracy between pipelines without and with DR for the same classifier. Source: the author.

	Test	Conover												Nemenyi			
	p-value adjustment technique	—				Bonferroni				Hommel				—			
	Classifier	KNN	MLP	SVM	RF	KNN	MLP	SVM	RF	KNN	MLP	SVM	RF	KNN	MLP	SVM	RF
Dimensionality	Projector																
2	PCA	0.0004	0.1836	0.6095	0.0121	0.5802	1.0000	1.0000	1.0000	0.3107	1.0000	1.0000	1.0000	0.1716	0.9000	0.9000	0.9000
	TSVD	0.0004	0.5035	0.7298	0.7845	0.5802	1.0000	1.0000	1.0000	0.3107	1.0000	1.0000	1.0000	0.1716	0.9000	0.9000	0.9000
	PCS	0.3388	0.5361	0.0459	0.7845	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	0.9000	0.9000	0.9000	0.9000
	KPCA	0.8403	0.2136	0.0394	0.8913	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	0.9000	0.9000	0.9000	0.9000
	UMAP	0.0004	0.8403	0.0725	0.7845	0.5802	1.0000	1.0000	1.0000	0.3107	1.0000	1.0000	1.0000	0.1716	0.9000	0.9000	0.9000
	Ivis	0.0004	1.0000	0.1789	0.7845	0.5802	1.0000	1.0000	1.0000	0.3107	1.0000	1.0000	1.0000	0.1716	0.9000	0.9000	0.9000
3	PCA	0.6045	0.0342	0.9141	0.1547	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	0.9000	0.9000	0.9000	0.9000
	TSVD	0.0004	0.1331	0.0380	0.0374	0.5802	1.0000	1.0000	1.0000	0.3107	1.0000	1.0000	1.0000	0.1716	0.9000	0.9000	0.9000
	PCS	0.0004	0.8403	0.0010	0.7845	0.5802	1.0000	1.0000	1.0000	0.3107	1.0000	0.6109	1.0000	0.1716	0.9000	0.3102	0.9000
	KPCA	0.0004	0.0205	0.1156	0.7845	0.5802	1.0000	1.0000	1.0000	0.3107	1.0000	1.0000	1.0000	0.1716	0.9000	0.9000	0.9000
	UMAP	0.0004	0.8403	0.0256	0.7845	0.5802	1.0000	1.0000	1.0000	0.3107	1.0000	1.0000	1.0000	0.1716	0.9000	0.9000	0.9000
	Ivis	0.1091	0.0000	0.3106	0.8913	1.0000	0.0254	1.0000	1.0000	1.0000	0.0179	1.0000	1.0000	0.9000	0.0098	0.9000	0.9000

Table 8 – Friedman *post-hoc* tests on validation accuracy between pipelines without and with DR for the same classifier. Source: the author.

	Test	Conover												Nemenyi			
	p-value adjustment technique	—				Bonferroni				Hommel				—			
	Classifier	KNN	MLP	SVM	RF	KNN	MLP	SVM	RF	KNN	MLP	SVM	RF	KNN	MLP	SVM	RF
Dimensionality	Projector																
2	PCA	0.3099	0.0033	0.0335	0.1253	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	0.9000	0.5691	0.9000	0.9000
	TSVD	0.7331	0.0391	0.0085	0.0027	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	0.9993	0.9000	0.9000	0.7969	0.5225
	PCS	0.8926	0.0583	0.0187	0.0000	1.0000	1.0000	1.0000	0.0060	1.0000	1.0000	1.0000	0.0055	0.9000	0.9000	0.9000	0.0017
	KPCA	0.1027	0.0049	0.0684	0.0167	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	0.9000	0.6623	0.9000	0.9000
	UMAP	0.0225	0.7654	0.0073	0.2868	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	0.9000	0.9000	0.7606	0.9000
	Ivis	0.0003	1.0000	0.3785	0.8926	0.4485	1.0000	1.0000	1.0000	0.2994	1.0000	1.0000	1.0000	0.1216	0.9000	0.9000	0.9000
3	PCA	0.1343	0.0000	0.0913	0.3785	1.0000	0.0606	1.0000	1.0000	1.0000	0.0499	1.0000	1.0000	0.9000	0.0186	0.9000	0.9000
	TSVD	0.1685	0.0020	0.1362	0.0252	1.0000	1.0000	1.0000	1.0000	1.0000	0.9967	1.0000	1.0000	0.9000	0.4577	0.9000	0.9000
	PCS	0.8590	0.7654	0.0000	0.0000	1.0000	1.0000	0.0036	0.0662	1.0000	1.0000	0.0034	0.0542	0.9000	0.9000	0.0010	0.0203
	KPCA	0.0969	0.0003	0.2275	0.0391	1.0000	0.4252	1.0000	1.0000	1.0000	0.2860	1.0000	1.0000	0.9000	0.1159	0.9000	0.9000
	UMAP	0.0183	0.8036	0.0000	0.0044	1.0000	1.0000	0.0080	1.0000	1.0000	1.0000	0.0074	1.0000	0.9000	0.9000	0.0023	0.6364
	Ivis	0.0007	0.0000	0.1752	0.0029	0.9300	0.0000	1.0000	1.0000	0.5532	0.0000	1.0000	1.0000	0.2185	0.0010	0.9000	0.5432

To illustrate how to interpret the cells of these tables, the first cell of Table 7 indicates that the unadjusted p-value yielded by the Conover *post-hoc* test between the baseline KNN pipeline with the one based on a two-dimensional PCA projection is 0.0004.

Looking at the adjusted p-values for train accuracy scores in Table 7, there are no comparisons with low likelihoods other than the one between MLP pipelines trained on the original set and on the three-dimensional Ivis projection. This low p-value, which was attested for both tests and both p-value adjustment techniques, is justified by the fact that the baseline MLP pipeline attained substantially lesser accuracy scores on average.

Regarding test accuracy scores, Table 8 has more occurrences of low p-values. Between the pipelines built on two-dimensional projections, only the PCS-based one recurrently attained low likelihoods. As such, the substantially low performance attained by the PCS projector for RF against the baseline pipeline is unlikely to have occurred by chance. For the pipelines built on three-dimensional projections, five p-values below 0.05 were attested by at least one test. All low probabilities involving MLP are due to its poor performance on the original data set. For the remaining comparisons with low likelihood involving SVM and RF, this means that the poorer performance of DR-encompassing pipelines when compared to the baselines is unlikely to have occurred by chance.

Overall, the majority of comparisons between pipelines without and with DR attained high probabilities assuming the null distribution to be true. This means that, for most comparisons, no evidence was found indicating substantial performance discrepancy.

For additional statistical significance analyses, Appendix B exposes the results of all comparisons involving different pipeline types.

4.1.3 Hyper-parameter inspection

Table 9 exposes the HPs that best optimized the objective function for each pipeline stratified by estimator. This table allows an analysis of how HP settings changed when the projection dimensionality went from two to three, as well as how importantly UMAP and Ivis weighted label information.

Overall, it seems that distinct HPs were attained for different combinations of projectors and classifiers. This might indicate that the HPO procedure did not converge or that the HPs with substantially discrepant values do not sensibly affect the resulting performance. Alternatively, this could mean merely that different estimator combinations required different HP settings to optimally perform.

Still, some preferences can be observed. One example is the unanimity of all Ivis estimators in picking `maaten` as their underlying pre-defined ANN. This aligns with its documentation, which recommends the `maaten` network for small data sets. Interestingly, Ivis estimators tended to higher values of `n_epochs_without_progress`, as well as UMAP

Table 9 – HPs of best-performing pipelines obtained by means of BO with 10-fold stratified CV for pipelines without and with DR. Source: the author.

		Classifier	KNN	MLP	SVM	RF
Dimensionality	Projector	Parameter				
50	—	Projector	—	—	—	—
		Classifier	leaf_size: 26 n_neighbors: 3 p: 1 weights: distance	activation: relu alpha: 0.7215564227807617 hidden_layer_sizes: (67, 54) max_iter: 2000	C: 20.41617597928149 kernel: linear max_iter: 100000 probability: True	max_depth: 5 n_estimators: 50 n_jobs: -1
2	PCA	Projector	—	—	—	—
		Classifier	leaf_size: 3 n_neighbors: 25 p: 5 weights: uniform	activation: relu alpha: 0.18737835928036276 hidden_layer_sizes: (82, 124) max_iter: 1194	C: 26.80380173673879 kernel: linear max_iter: 100000 probability: True	max_depth: 3 n_estimators: 135 n_jobs: -1
	TSVD	Projector	algorithm: arpack	algorithm: arpack	algorithm: arpack	algorithm: arpack
		Classifier	leaf_size: 17 n_neighbors: 25 p: 1 weights: distance	activation: tanh alpha: 0.0001 hidden_layer_sizes: (85, 106) max_iter: 479	C: 7.1205395004936 kernel: linear max_iter: 100000 probability: True	max_depth: 12 n_estimators: 113 n_jobs: -1
	PCS	Projector	—	—	—	—
		Classifier	leaf_size: 5 n_neighbors: 19 p: 5 weights: uniform	activation: tanh alpha: 7.574213595095741 hidden_layer_sizes: (93, 109) max_iter: 100	C: 543.3626102483709 gamma: 16.892503181802894 kernel: rbf max_iter: 100000 probability: True	max_depth: 3 n_estimators: 150 n_jobs: -1
	KPCA	Projector	eigen_solver: dense gamma: 0.7631660469278523 kernel: sigmoid n_jobs: -1	eigen_solver: dense kernel: linear n_jobs: -1	eigen_solver: dense gamma: 0.5129059051480938 kernel: linear n_jobs: -1	eigen_solver: dense gamma: 0.1168052744110393 kernel: sigmoid n_jobs: -1
		Classifier	leaf_size: 21 n_neighbors: 16 p: 4 weights: distance	activation: relu alpha: 0.0001 hidden_layer_sizes: (90, 50) max_iter: 2000	C: 127.65749874708352 kernel: linear max_iter: 100000 probability: True	max_depth: 11 n_estimators: 50 n_jobs: -1
	UMAP	Projector	n_epochs: 1694 n_neighbors: 3 target_weight: 0.6887647298354286	n_epochs: 282 n_neighbors: 15 target_weight: 0.6916103701632366	n_epochs: 2000 n_neighbors: 3 target_weight: 0.2387735663333006	n_epochs: 2000 n_neighbors: 3 target_weight: 0.7879949317192984
		Classifier	leaf_size: 15 n_neighbors: 4 p: 3 weights: distance	activation: tanh alpha: 0.0001 hidden_layer_sizes: (69, 50) max_iter: 100	C: 985.2902676692498 gamma: 56.595855465395 06 kernel: rbf max_iter: 100000 probability: True	max_depth: 3 n_estimators: 50 n_jobs: -1

Continued on next page

Table 9 – HPs of best-performing pipelines obtained by means of BO with 10-fold stratified CV for pipelines without and with DR. Source: the author.

		Classifier	KNN	MLP	SVM	RF
Dimensionality	Projector	Parameter				
	Ivis	Projector	k: 103 model: hinton n_epochs_without_progress: 50 supervision_weight: 0.8013832491892193	k: 38 model: hinton n_epochs_without_progress: 40 supervision_weight: 1.0	k: 3 model: hinton n_epochs_without_progress: 30 supervision_weight: 1.0	k: 29 model: hinton n_epochs_without_progress: 50 supervision_weight: 1.0
		Classifier	leaf_size: 3 n_neighbors: 30 p: 1 weights: distance	activation: tanh alpha: 0.0001 hidden_layer_sizes: (127, 150) max_iter: 2000	C: 447.8756201510728 gamma: 0.01 kernel: rbf max_iter: 100000 probability: True	max_depth: 14 n_estimators: 150 n_jobs: -1
3	PCA	Projector	—	—	—	—
		Classifier	leaf_size: 29 n_neighbors: 19 p: 1 weights: distance	activation: tanh alpha: 0.0001 hidden_layer_sizes: (55, 50) max_iter: 100	C: 331.3969277503116 gamma: 6.830188552360055 kernel: rbf max_iter: 100000 probability: True	max_depth: 8 n_estimators: 62 n_jobs: -1
	TSVD	Projector	algorithm: arpack	algorithm: arpack	algorithm: arpack	algorithm: arpack
		Classifier	leaf_size: 4 n_neighbors: 19 p: 3 weights: distance	activation: relu alpha: 0.6779758749165763 hidden_layer_sizes: (150, 50) max_iter: 100	C: 629.9676083781482 gamma: 0.01 kernel: rbf max_iter: 100000 probability: True	max_depth: 15 n_estimators: 127 n_jobs: -1
	PCS	Projector	—	—	—	—
		Classifier	leaf_size: 13 n_neighbors: 6 p: 1 weights: distance	activation: tanh alpha: 1.733311249387322 hidden_layer_sizes: (141, 135) max_iter: 799	C: 229.59492508889957 gamma: 1.0172937489968852 kernel: rbf max_iter: 100000 probability: True	max_depth: 10 n_estimators: 60 n_jobs: -1
	KPCA	Projector	eigen_solver: dense kernel: linear n_jobs: -1	eigen_solver: dense gamma: 0.19497722737592268 kernel: sigmoid n_jobs: -1	eigen_solver: dense gamma: 0.31690871193087605 kernel: sigmoid n_jobs: -1	eigen_solver: dense kernel: linear n_jobs: -1
		Classifier	leaf_size: 29 n_neighbors: 17 p: 1 weights: distance	activation: tanh alpha: 0.0001 hidden_layer_sizes: (82, 50) max_iter: 2000	C: 794.869407756802 gamma: 0.01 kernel: sigmoid max_iter: 100000 probability: True	max_depth: 8 n_estimators: 50 n_jobs: -1
	UMAP	Projector	n_epochs: 2000 n_neighbors: 3 target_weight: 0.0	n_epochs: 1852 n_neighbors: 25 target_weight: 0.0	n_epochs: 960 n_neighbors: 3 target_weight: 0.8473658289906908	n_epochs: 1987 n_neighbors: 4 target_weight: 0.0

Continued on next page

Table 9 – HPs of best-performing pipelines obtained by means of BO with 10-fold stratified CV for pipelines without and with DR. Source: the author.

		Classifier	KNN	MLP	SVM	RF
Dimensionality	Projector	Parameter				
		Classifier	leaf_size: 6 n_neighbors: 3 p: 5 weights: uniform	activation: tanh alpha: 8.46485306798708 hidden_layer_sizes: (126, 114) max_iter: 1552	C: 255.06009716561346 gamma: 64.4382973678162 kernel: rbf max_iter: 100000 probability: True	max_depth: 4 n_estimators: 82 n_jobs: -1
	Ivis	Projector	k: 28 model: hinton n_epochs_without_progress: 50 supervision_weight: 1.0	k: 96 model: hinton n_epochs_without_progress: 50 supervision_weight: 0.9159628556325711	k: 103 model: hinton n_epochs_without_progress: 50 supervision_weight: 1.0	k: 3 model: hinton n_epochs_without_progress: 30 supervision_weight: 0.7449338390496758
		Classifier	leaf_size: 5 n_neighbors: 3 p: 2 weights: distance	activation: tanh alpha: 0.3990511961539001 hidden_layer_sizes: (50, 68) max_iter: 1240	C: 898.4164364790793 gamma: 0.01 kernel: rbf max_iter: 100000 probability: True	max_depth: 15 n_estimators: 134 n_jobs: -1

for the most part with its `n_epochs` HP, thus allowing more iterations before the learning procedure stops. Regarding the weight of the labels on the learning, Ivis heavily relied on supervised learning for all pipelines. Curiously, UMAP showed a distinct behavior in this regard by relying more on label information for two-dimensional projections and disregarding it entirely for most of the three-dimensional ones.

4.2 Independent testing

The analyses of the products of the independent testing phase are divided in two parts: visual inspection (Section 4.2.1) and prediction quality measurements (Section 4.2.2). As mentioned in Chapter 3, the results of this phase are distinct from the previous one for they are obtained by exposing the produced pipelines to a holdout set that was unseen thus far for projection and classification.

4.2.1 Visual inspection

Figure 7 portrays two-dimensional representations of the data that the projectors yielded and the classifiers used to predict the HIA of small molecules. All but the UMAP projections seem to agree on a representation that somewhat resembles a diagonal line with substantial inter-cluster overlap. Among them, some projectors, such as TSVD and PCS, tend to produce more spread-out visualizations than others, such as PCS and KPCA.

The considered selection of linear techniques, which comprises PCA, TSVD, and PCS, produce identical representations across pipelines, for they are devoid of HPs and stochastic procedures. Even KPCA projections should not vary much since its `eigen_solver` HP was set to use a deterministic routine and its other HPs remained relatively stable overall. In contrast, UMAP and Ivis representations are expected to vary more across pipelines, as they employ iterative optimization processes influenced by multiple HPs. Still, Ivis behaved similarly to the other techniques, leaving UMAP projections as the most different-looking ones.

By jointly inspecting all projections of the train set, it is noticeable that HIA (+) molecules seem to form a denser cluster than the HIA (–) ones on most representations. Despite the attempts of UMAP to depict the data differently, the representations ended up comprising heterogeneous clusters with substantial inter-class overlap.

Figure 8 seem to maintain the observed tendencies of Figure 7. The most noticeable difference lies in the cluster densities, something that is attributable to the fact that the test set is substantially smaller than the train one (which was also used for validation purposes, as explained in Chapter 3). Structural patterns observed in the mappings of the train set seem to be retained in the predicted projections of the test set.

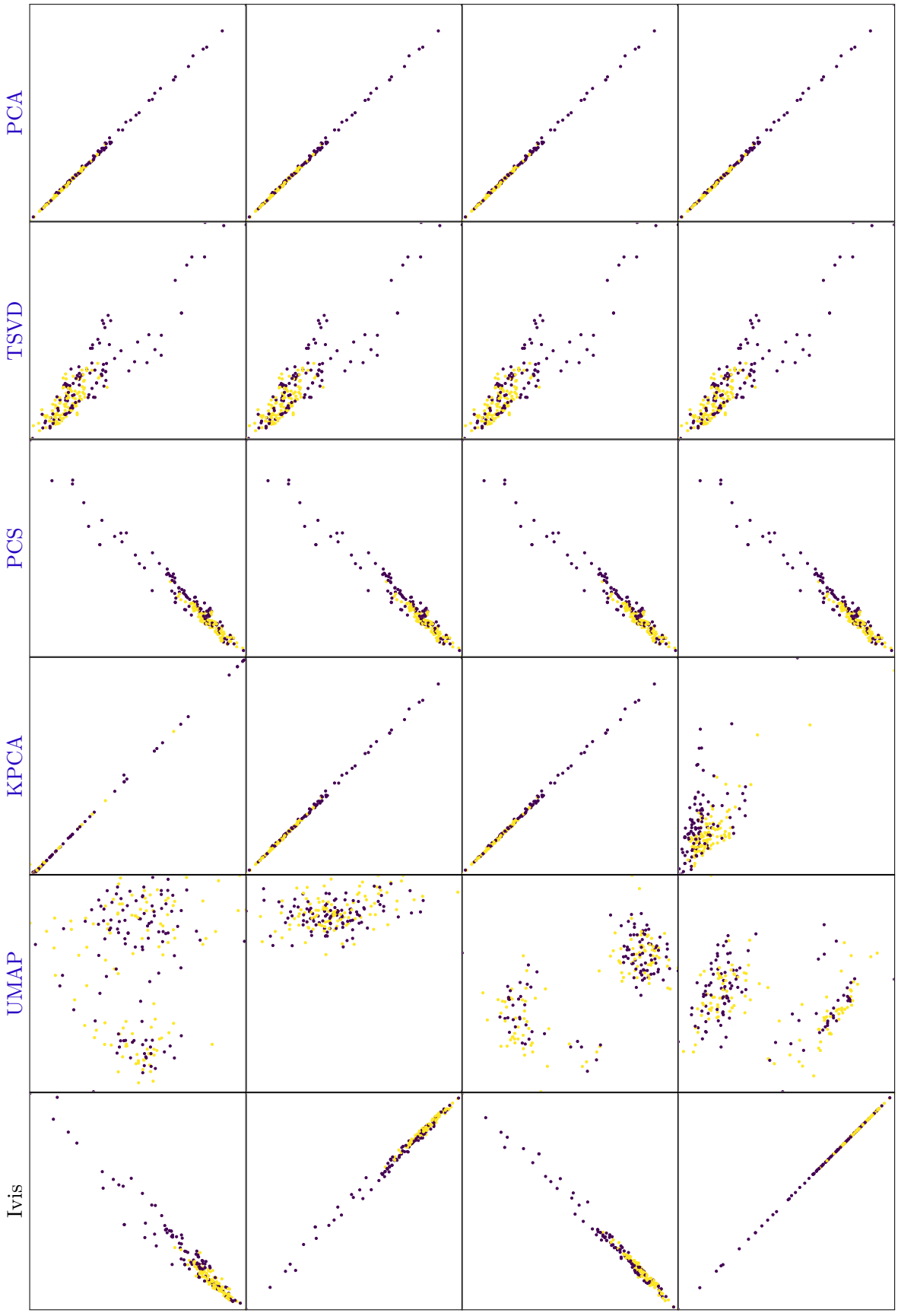


Figure 7 – Two-dimensional projections of the train set for all classifiers, with HIA (–) samples represented in purple and HIA (+) in yellow. Source: the author.

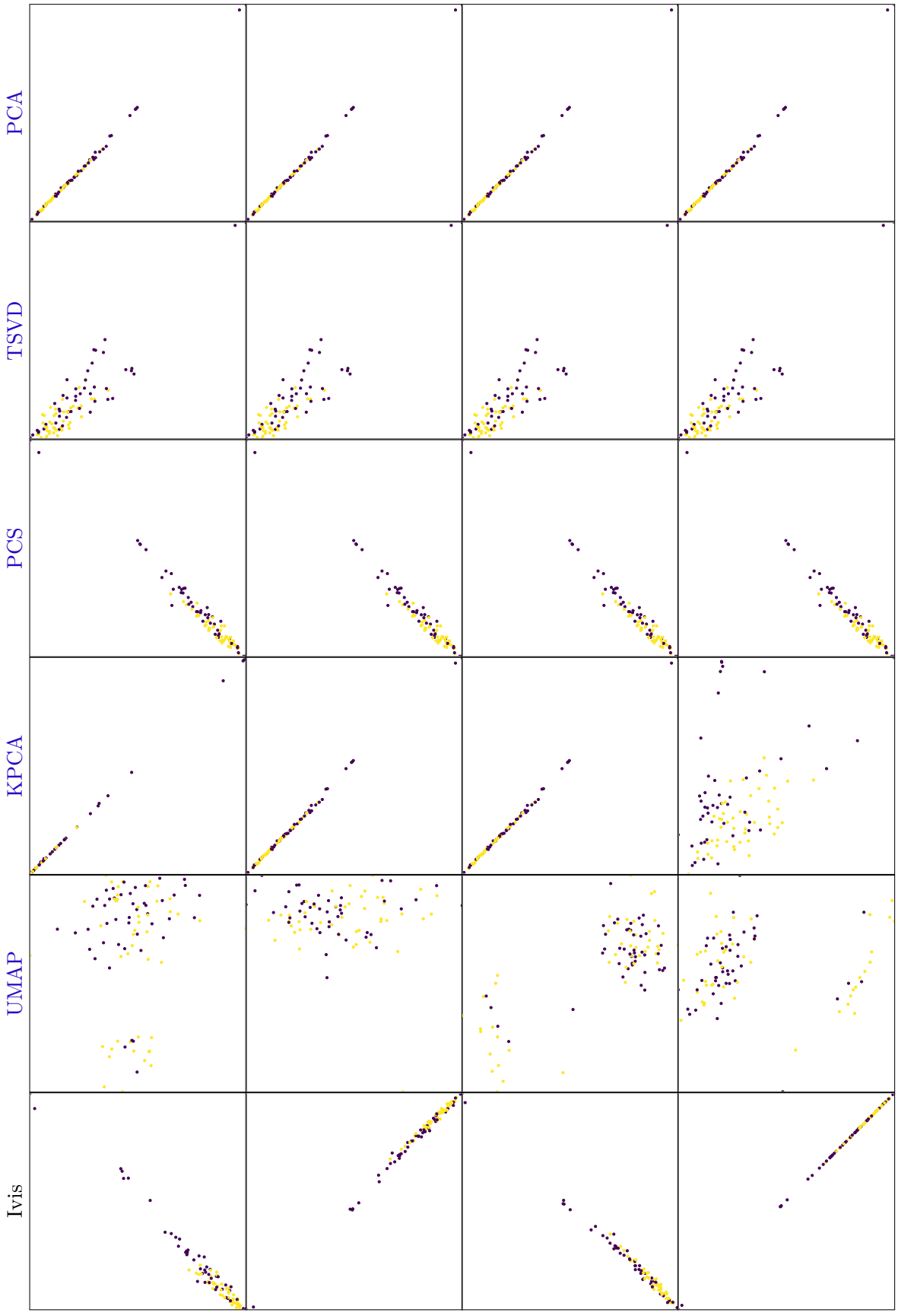
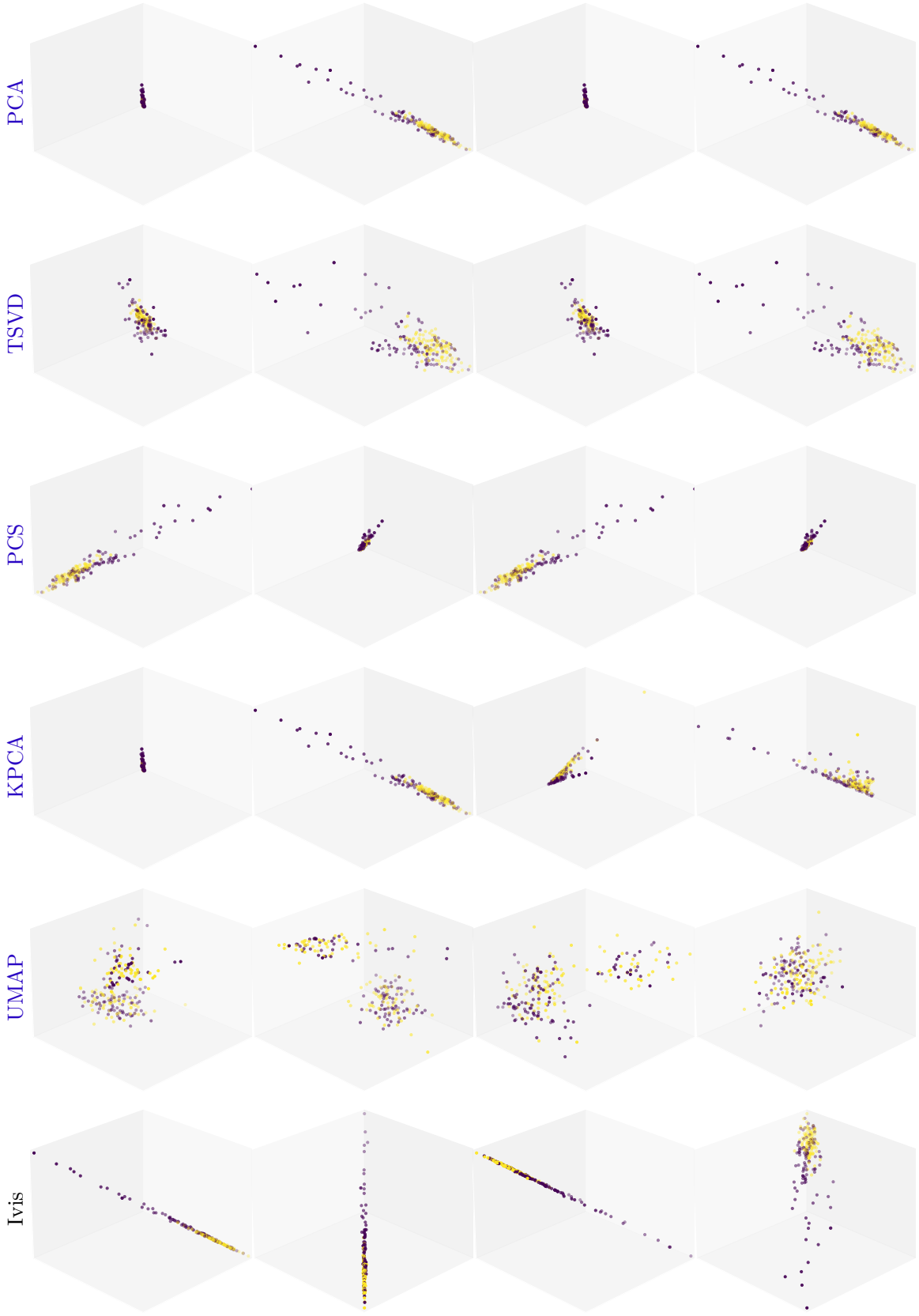
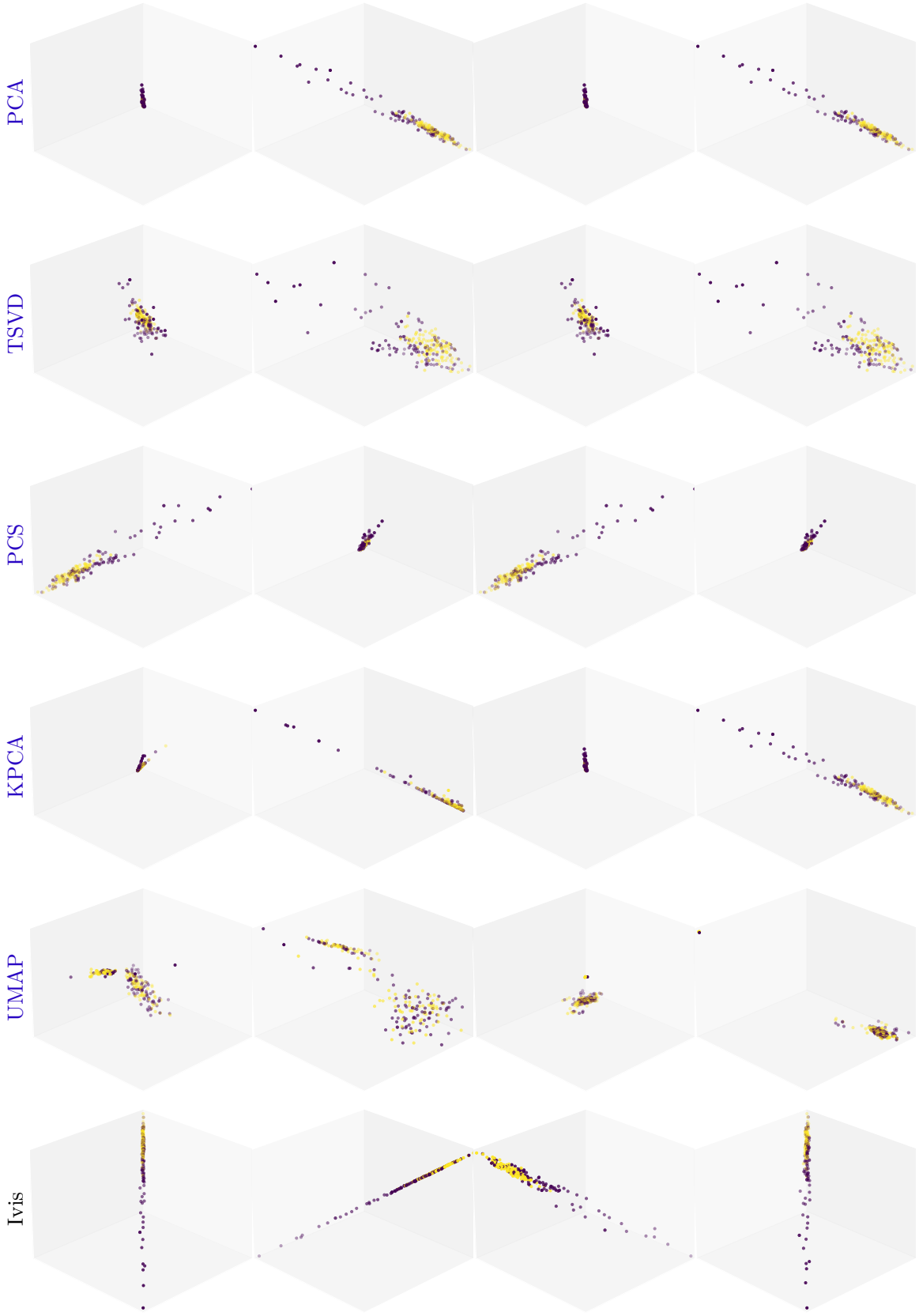


Figure 8 – Two-dimensional projections of the test set for all classifiers, with HIA (–) samples represented in purple and HIA (+) in yellow. Source: the author.



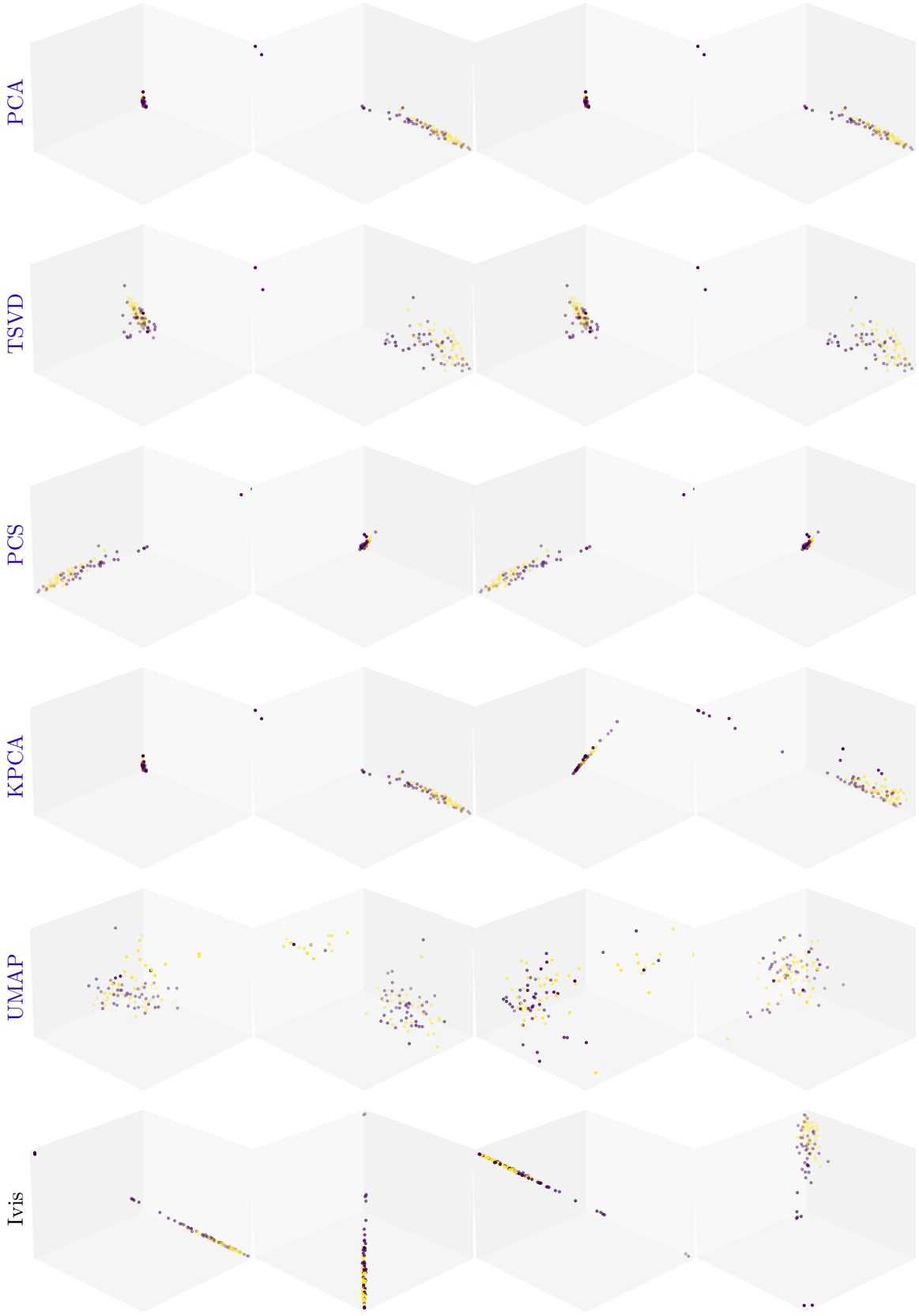
(a) KNN (b) MLP

Figure 9 – Two perspectives of three-dimensional projections of the train set for KNN and MLP, with HIA (–) samples represented in purple and HIA (+) in yellow. Source: the author.



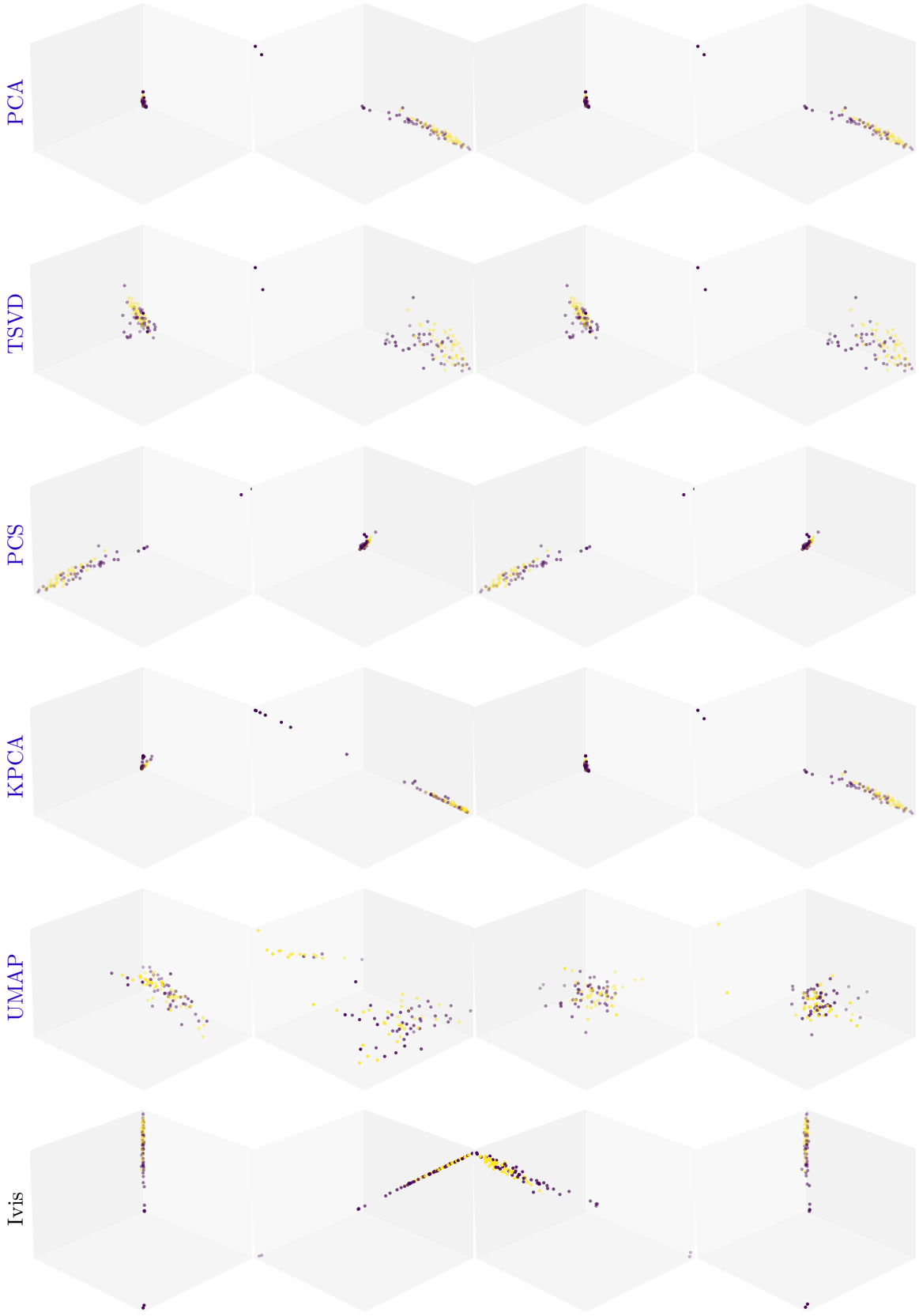
(a) SVM (b) RF

Figure 10 – Two perspectives of three-dimensional projections of the train set for SVM and RF, with HIA (–) samples represented in purple and HIA (+) in yellow. Source: the author.



(a) KNN (b) MLP

Figure 11 – Two perspectives of three-dimensional projections of the test set for KNN and MLP, with HIA (–) samples represented in purple and HIA (+) in yellow. Source: the author.



(a) SVM (b) RF

Figure 12 – Two perspectives of three-dimensional projections of the test set for SVM and RF, with HIA (–) samples represented in purple and HIA (+) in yellow. Source: the author.

Adding another feature to the projected space might enrich the projections, as projectors will have another dimension to use when representing the information. In contrast, this might represent a hardening of the optimization problem to be solved by these algorithms, especially if they perform iterative optimization processes. Figures 9 and 10 depict three-dimensional projections of the train set, whereas Figures 11 and 12 portray three-dimensional representations of the test set. Two perspectives of the projected space were taken for each dimension to compensate for the lack of means to rotate and interact with the three-dimensional representations.

Overall, the three-dimensional representations maintain spatial arrangements observed in Figures 7 and 8: most techniques mapped the data such that the resulting structure resembles a diagonal line. Likewise, UMAP depicted the data as sparse, heterogeneous clusters. Although it managed to spread the data throughout the space, it does not seem that this effort resulted in an inter-cluster separation that is any better than the one observed in the remaining projections.

For the sake of completeness, Appendix C concatenates the projections of the train and test sets, allowing for a visualization of the entire data set via two- and three-dimensional projections. Even though this compiled visualization mixes representations of instances seen in the learning process with unforeseen ones whose positioning was predicted, these visualizations might be of value to see the entirety of the data used in training and testing.

4.2.2 Prediction quality measurements

Table 10 describes the performance of the pipelines with the best HP settings found in the pipeline tuning phase. Similar to what was observed with the pipeline tuning results, some MLP estimators stand out due to the fact that their accuracy measurements hover around 50%. In all fairness, only two MLP pipelines showed this behavior here, less than half of the five that did so on the previous phase. Nevertheless, this issue should be clarified, something that is fortunately facilitated by the availability of additional measurements.

Noticeably for these bad-performing, MLP-based pipelines, either their sensitivity or their specificity scores are set to very discrepant values, leading to the understanding that, for these pipelines, one class is almost always predicted. These problematic pipelines can also be detected by the F1 score and the AUC due to the discrepancies of sensitivity and specificity. On the bright side, all pipelines without DR seem to be behaving normally in this phase, allowing us to use them as baselines for the ones with DR.

Accuracy-wise, the baseline SVM and RF pipelines are the best-performing ones. Their other measurements are also competitive. In terms of overall sensibility and specificity measurements, all models seem to better predict HIA (+) molecules than HIA (–) ones.

Table 10 – Independent test results for the best pipelines found. Source: the author.

		Measure	Accuracy	F1	AUC	MCC	Sensitivity	Specificity
Dimensionality	Projector	Classifier						
50	—	KNN	0.706522	0.727273	0.706522	0.417911	0.782609	0.630435
		MLP	0.717391	0.750000	0.717391	0.450377	0.847826	0.586957
		SVM	0.793478	0.804124	0.793478	0.590455	0.847826	0.739130
		RF	0.804348	0.820000	0.804348	0.618115	0.891304	0.717391
2	PCA	KNN	0.706522	0.747664	0.706522	0.436926	0.869565	0.543478
		MLP	0.717391	0.754717	0.717391	0.456435	0.869565	0.565217
		SVM	0.706522	0.756757	0.706522	0.453539	0.913043	0.500000
		RF	0.728261	0.761905	0.728261	0.475923	0.869565	0.586957
	TSVD	KNN	0.739130	0.764706	0.739130	0.489979	0.847826	0.630435
		MLP	0.663043	0.699029	0.663043	0.335830	0.782609	0.543478
		SVM	0.673913	0.722222	0.673913	0.370991	0.847826	0.500000
		RF	0.760870	0.775510	0.760870	0.526235	0.826087	0.695652
	PCS	KNN	0.695652	0.720000	0.695652	0.397360	0.782609	0.608696
		MLP	0.521739	0.676471	0.521739	0.149071	1.000000	0.043478
		SVM	0.684783	0.718447	0.684783	0.380608	0.804348	0.565217
		RF	0.706522	0.715789	0.706522	0.413925	0.739130	0.673913
	KPCA	KNN	0.717391	0.729167	0.717391	0.436436	0.760870	0.673913
		MLP	0.684783	0.707071	0.684783	0.373920	0.760870	0.608696
		SVM	0.706522	0.756757	0.706522	0.453539	0.913043	0.500000
		RF	0.750000	0.767677	0.750000	0.505892	0.826087	0.673913
	UMAP	KNN	0.728261	0.742268	0.728261	0.459243	0.782609	0.673913
		MLP	0.717391	0.763636	0.717391	0.472456	0.913043	0.521739
		SVM	0.739130	0.750000	0.739130	0.480079	0.782609	0.695652
		RF	0.728261	0.747475	0.728261	0.461901	0.804348	0.652174
	Ivis	KNN	0.750000	0.785047	0.750000	0.528910	0.913043	0.586957
		MLP	0.739130	0.760000	0.739130	0.485662	0.826087	0.652174
		SVM	0.760870	0.792453	0.760870	0.547723	0.913043	0.608696
		RF	0.739130	0.750000	0.739130	0.480079	0.782609	0.695652
3	PCA	KNN	0.706522	0.752294	0.706522	0.444513	0.891304	0.521739
		MLP	0.717391	0.767857	0.717391	0.482805	0.934783	0.500000
		SVM	0.684783	0.743363	0.684783	0.415376	0.913043	0.456522
		RF	0.728261	0.747475	0.728261	0.461901	0.804348	0.652174
	TSVD	KNN	0.673913	0.727273	0.673913	0.377964	0.869565	0.478261
		MLP	0.663043	0.715596	0.663043	0.350931	0.847826	0.478261
		SVM	0.673913	0.732143	0.673913	0.386244	0.891304	0.456522
		RF	0.684783	0.712871	0.684783	0.376848	0.782609	0.586957
	PCS	KNN	0.641304	0.685714	0.641304	0.294619	0.782609	0.500000
		MLP	0.500000	0.666667	0.500000	0.000000	1.000000	0.000000
		SVM	0.641304	0.727273	0.641304	0.364073	0.956522	0.326087
		RF	0.717391	0.754717	0.717391	0.456435	0.869565	0.565217
	KPCA	KNN	0.695652	0.735849	0.695652	0.410792	0.847826	0.543478
		MLP	0.695652	0.730769	0.695652	0.405340	0.826087	0.565217
		SVM	0.706522	0.737864	0.706522	0.425385	0.826087	0.586957
		RF	0.728261	0.752475	0.728261	0.465519	0.826087	0.630435
	UMAP	KNN	0.717391	0.734694	0.717391	0.438529	0.782609	0.652174
		MLP	0.500000	0.000000	0.500000	0.000000	0.000000	1.000000
		SVM	0.750000	0.767677	0.750000	0.505892	0.826087	0.673913

Continued on next page

Table 10 – Independent test results for the best models found. Source: the author.

		Measure	Accuracy	F1	AUC	MCC	Sensitivity	Specificity
Dimensionality	Projector	Classifier						
		RF	0.663043	0.693069	0.663043	0.332513	0.760870	0.565217
	Ivis	KNN	0.728261	0.742268	0.728261	0.459243	0.782609	0.673913
		MLP	0.706522	0.737864	0.706522	0.425385	0.826087	0.586957
		SVM	0.706522	0.742857	0.706522	0.430597	0.847826	0.565217
		RF	0.750000	0.767677	0.750000	0.505892	0.826087	0.673913

Akin to what was done when inspecting the results of the previous phase, the measurements of Table 10 were portrayed in individual plots stratifying first by classifier instead of by projector, thus allowing relative comparisons between DR-encompassing pipelines and their respective baseline counterparts.

Figure 13 portray accuracy measurements. It is perceptible that KNN pipelines using UMAP and Ivis surpassed the KNN baseline model. Ivis also scored competitively for MLP, surpassing the baseline here as well. The UMAP model for two dimensions scored as good as the baseline MLP model, although the same cannot be said about the one for three dimensions. Conversely, for SVM and RF, although competitive, no pipelines with DR surpassed the baseline models.

The tendencies are similar for the F1 scores displayed in Figure 14, with some

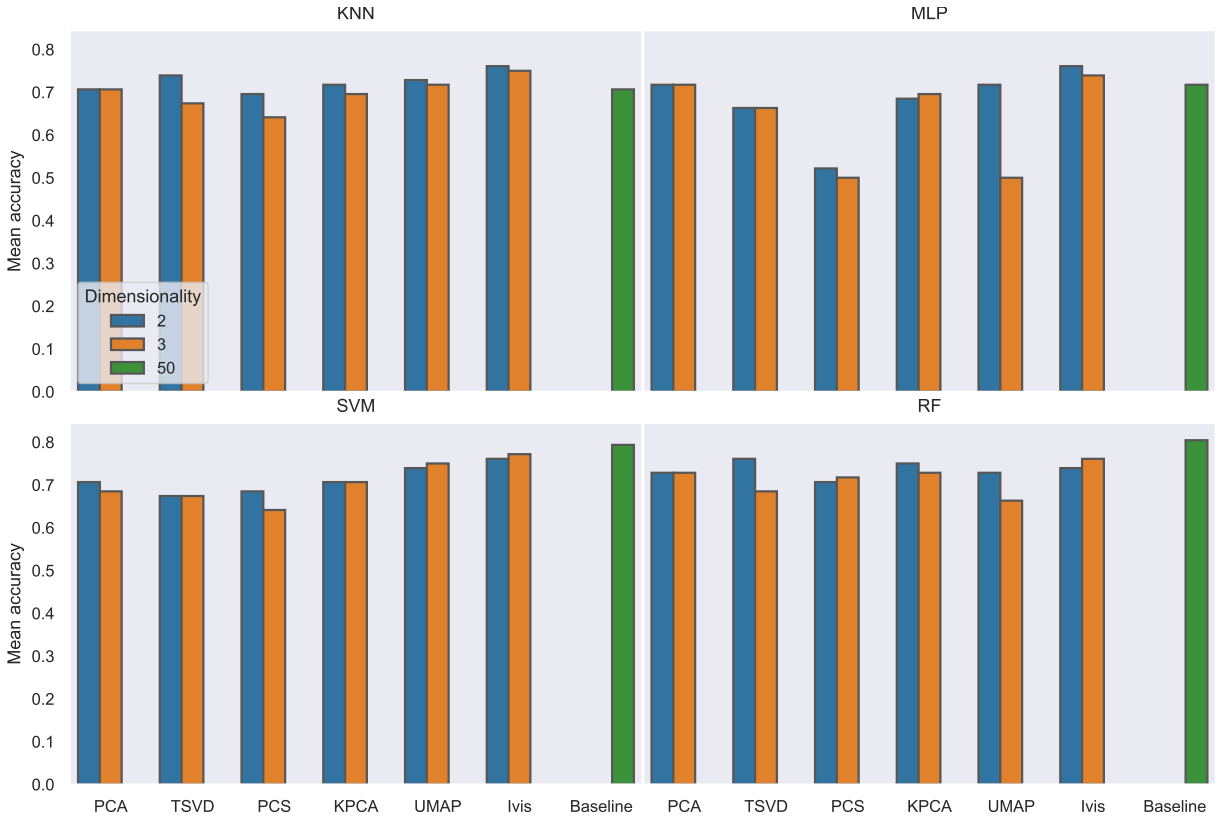


Figure 13 – Independent test accuracy measures for each classifier grouped by projector and stratified by dimensionality. Source: the author.

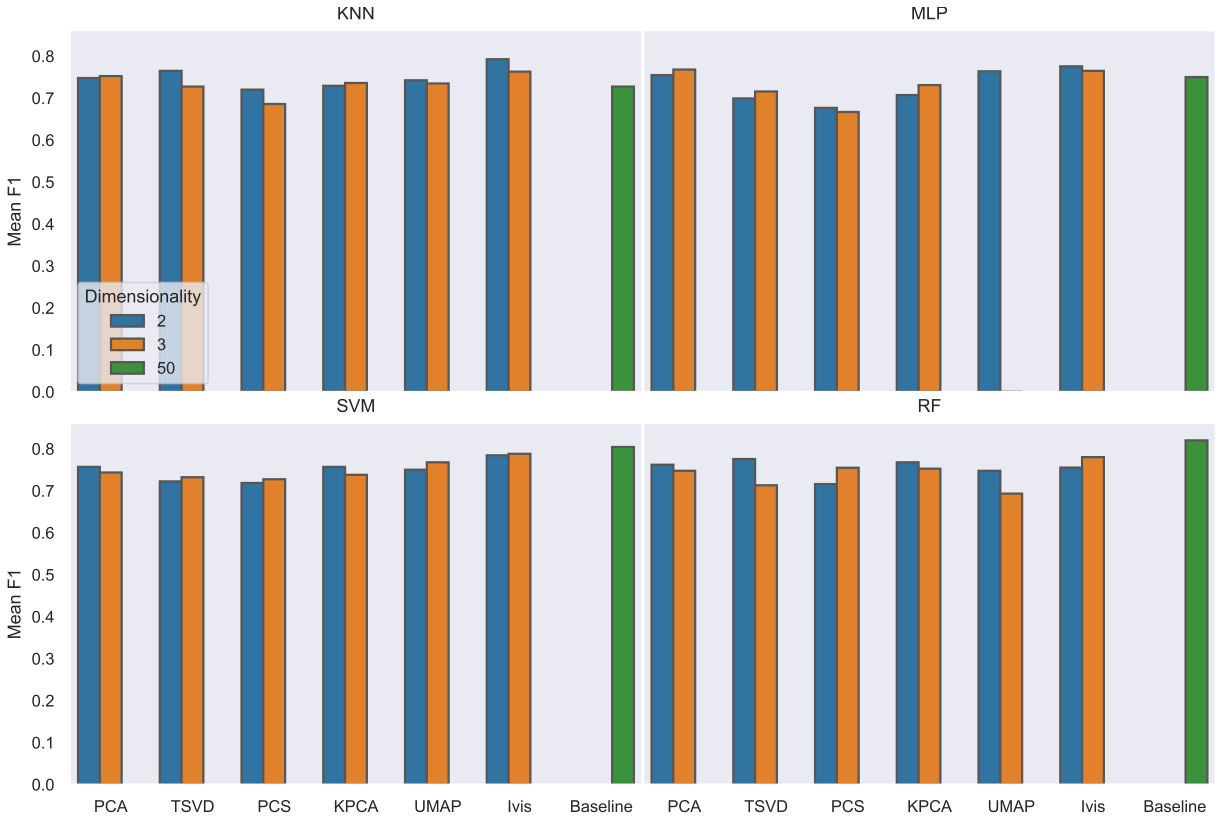


Figure 14 – Independent test F1 measures for each classifier grouped by projector and stratified by dimensionality. Source: the author.

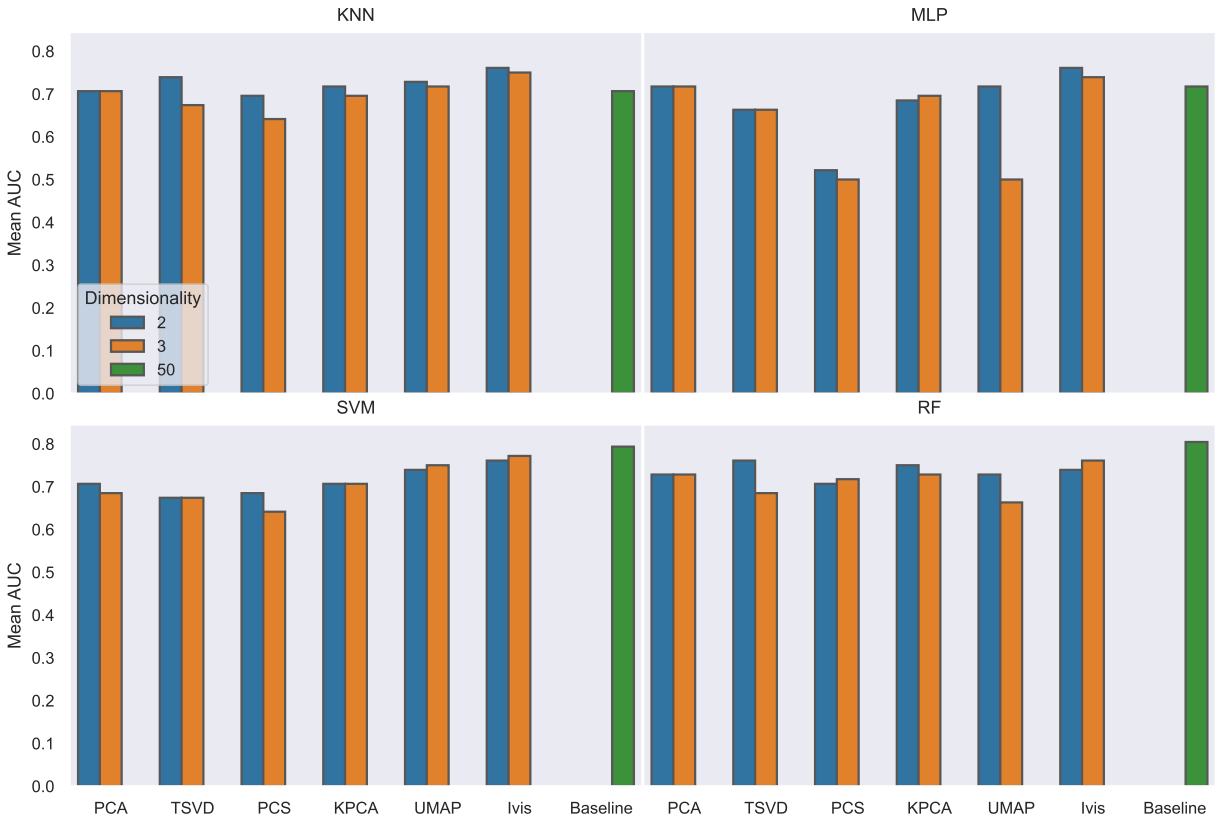


Figure 15 – Independent test AUC measures for each classifier grouped by projector and stratified by dimensionality. Source: the author.

DR-encompassing pipelines surpassing the baseline ones for KNN and MLP, but none for SVM and RF. Notably, the score obtained for the pipeline based on a three-dimensional UMAP projection and a MLP classifier scored zero, indicating that this pipeline is one of the MLP-based ones that did not adequately labeled the molecules of the hold out set.

Figure 15 facilitates the visual inspection of AUC measurements. A similar arrangement is observed overall to Figure 13, including the MLP models that ended up predicting only one class for all instances of the test set, consequently scoring values near 50%. At this point of the analysis, it has become evident that some MLP-based pipelines, despite the extensive HPO performed, are experiencing difficulties in properly labeling molecules.

The MCC values shown in Figure 16 show a visually distinct disposition when compared to the previous figures. Nevertheless, the tendency is ultimately the same: pipelines with DR winning for KNN and MLP, but not for SVM and RF. In fact, for the last two classifiers, the gap between the DR-encompassing pipelines and the baseline ones is wider here, favoring the latter. Again, the bad-performing MLP models are easily distinguishable due to their low or null scores, further establishing the understanding that they are ultimately no better than a random model.

Lastly, Figures 17 and 18 expose the obtained sensitivity and specificity scores, respectively. Exploring the information portrayed in these figures facilitates the strati-

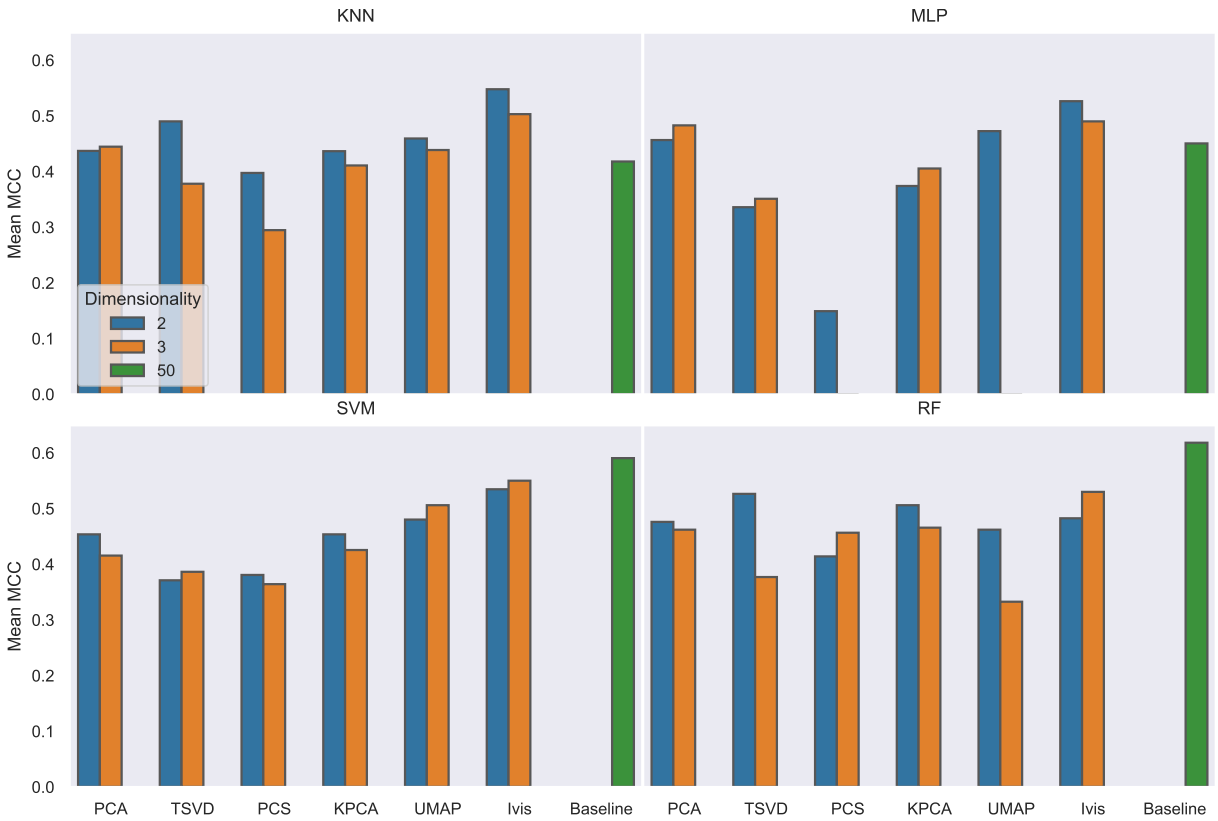


Figure 16 – Independent test MCC measures for each classifier grouped by projector and stratified by dimensionality. Source: the author.

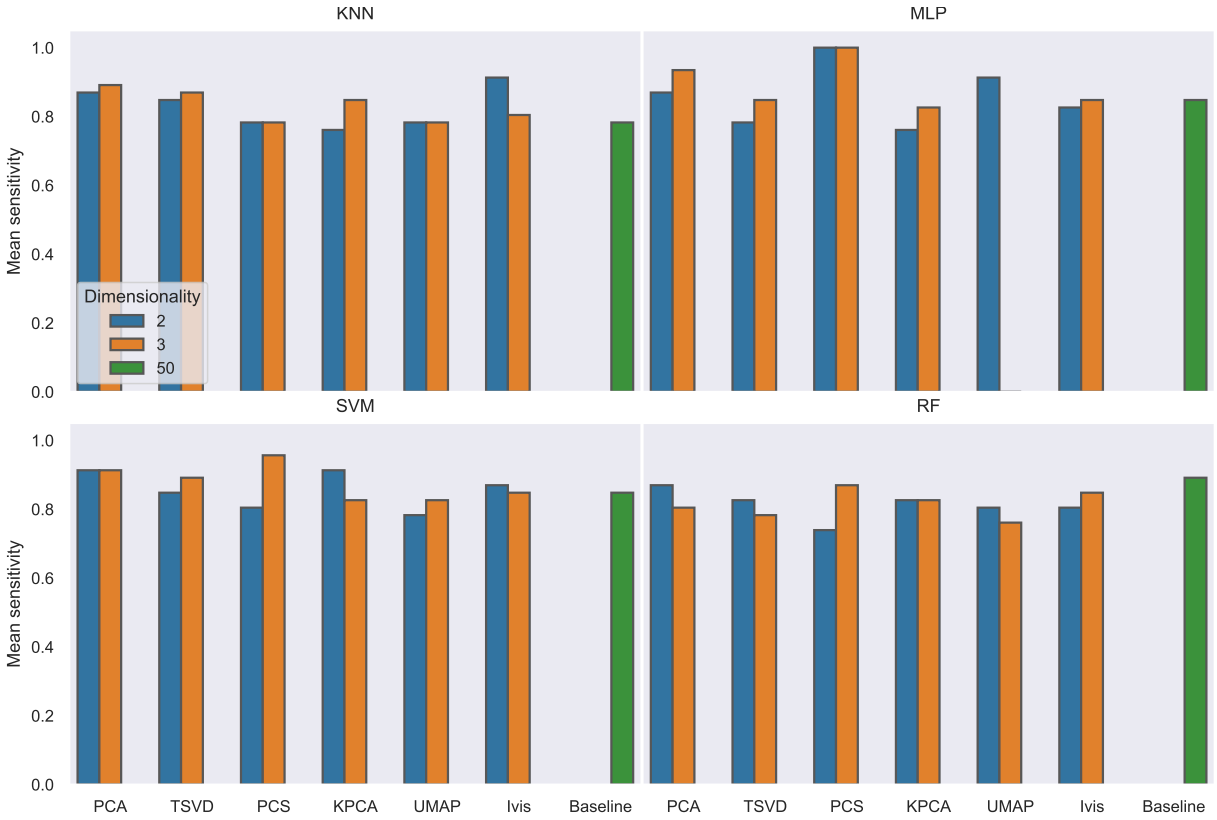


Figure 17 – Independent test sensitivity measures for each classifier grouped by projector and stratified by dimensionality. Source: the author.

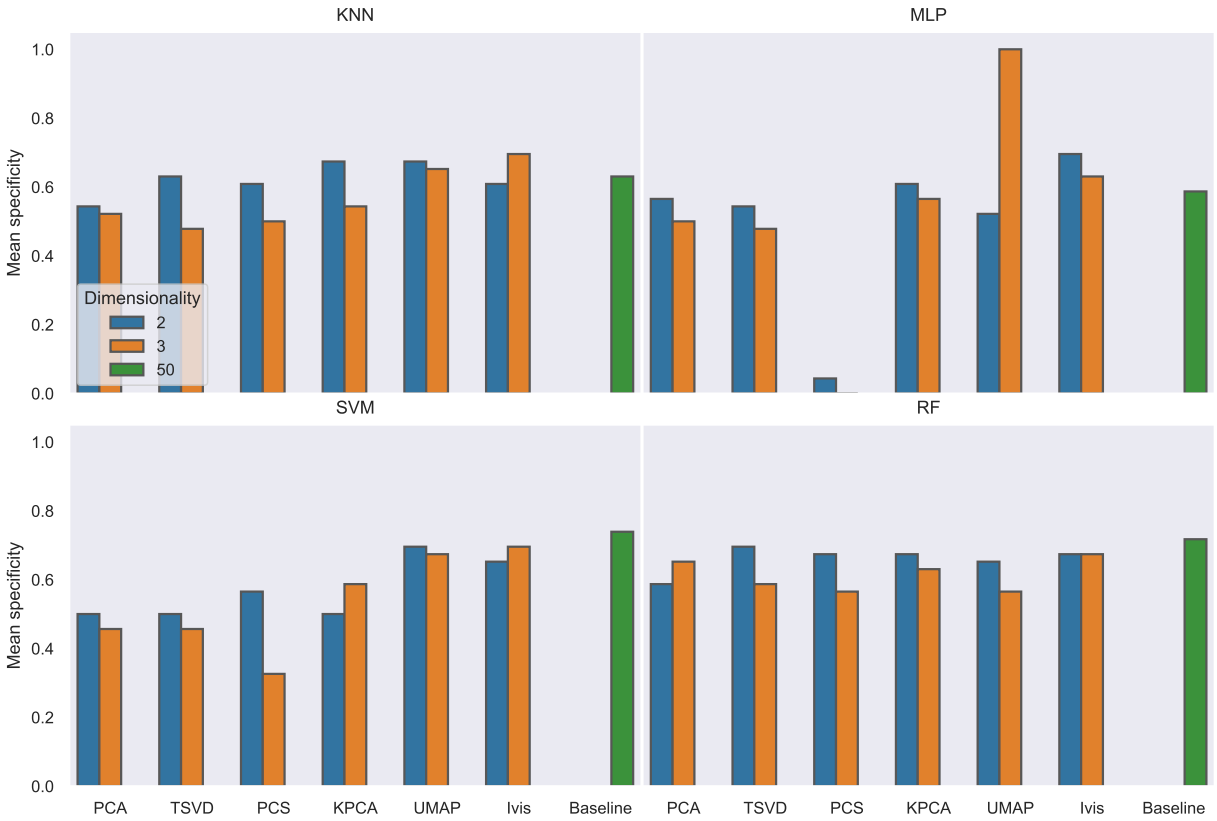


Figure 18 – Independent test specificity measures for each classifier grouped by projector and stratified by dimensionality. Source: the author.

fication of the worst-performing MLP models between those that only predict HIA (+) (i.e., those with high sensitivity and low specificity) from those that only issue HIA (–) labels (i.e., those with low sensitivity and high specificity). Furthermore, comparing these figures enables a visual discernment on how the produced pipelines, on average, tend to correctly predict HIA (+) molecules more often than HIA (–) ones. Other than that, the overarching disposition observed in them is similar to the previous figures.

CHAPTER

5

FINAL REMARKS

The overarching goal of this work was to determine whether applying DR via feature extraction on physicochemical and structural properties of small molecules would be feasible for visual inspection and HIA prediction. In anticipation of this, the introductory chapter (Chapter 1) allowed the reader to get acquainted with aspects of the computational, pharmaceutical, and chemical realms related to the problem being tackled, viz., drug formulation design and changes brought about by Big Data on the KDD process. Subsequently, a theoretical chapter (Chapter 2) further elaborated on concepts of these realms by clarifying how HIA *de facto* occurs and what molecular descriptors are, as well as framing projection and classification as tasks of ML, detailing means of carrying out HPO, and elaborating on statistical significance analyses.

The construction made by these chapters led to Chapter 3, wherein the design and *modus operandi* of the experiments conducted by this work were defined and explained. Its main product is a pipeline structure that begins with 304 molecular representations in SMILES strings—which are described in terms of eight physicochemical and forty-two structural properties—and ends with trained and validated projectors and classifiers. Ultimately, it is possible to generate two- and three-dimensional projections of the small molecules, as well as predictions of their permeability in the epithelial barrier of the GI tract. The devised pipeline structure allows the practitioner to switch projectors and classifiers as drop-in replacements, thus adapting the pipeline to the idiosyncrasies of the problem and data set at hand.

To tackle the problem under study, a representative selection of six projectors (Ivis, KPCA, PCA, PCS, TSVD, and UMAP) and four classifiers (KNN, MLP, RF, and SVM) was made. All possible projector-classifier combinations were tested for two- and three-dimensional projected spaces, resulting in forty-eight pipelines that employed some DR technique. Four pipelines devoid of DR were created, amounting to fifty-two pipelines overall to have a baseline against which to compare the DR-encompassing pipelines.

As part of the methodology employed to generate trained and validated pipelines, each pipeline was individually fine-tuned with BO coupled with 10-fold CV, thus incurring in 520,000 HP evaluations in total. Measurements of this pipeline tuning phase were collected for inspection and statistical significance analysis, as well as the best HP settings found. Subsequently, the trained and validated pipelines underwent an independent test with a hold-out set of small molecules. The projections of the train and test sets produced by each pipeline were collected for visual inspection and the performance of the models is quantified with six measures, namely accuracy, F1 score, AUC, MCC, sensitivity, and specificity. Analyzing these results, which were exposed and discussed in Chapter 4, helped paint a comprehensive picture of the impacts of adding DR techniques both in terms of visual-exploratory analyses of small molecule data and HIA prediction.

Drawing from interpretations of the results exposed in the previous chapter and grounded by the contextualization provided by the earlier ones, this chapter closes this work by drawing some conclusions (Section 5.1) and enumerating future work (Section 5.2).

5.1 Conclusions

As previously stated, applying DR via feature extraction can be characterized as a means of lossy compression, as it results in a representation of the original data set that is described in a smaller vector space. As such, some quality loss is expected. Unlike feature selection techniques, however, feature extraction techniques can, with their respective learning procedures, internalize a mapping that produces projections with desirable properties, such as simpler feature spaces and greater inter-cluster separation, potentially compensating the compression effects. The mere reduction of the feature space might suffice for classification algorithms to improve their performance, as the curse of dimensionality is mitigated. The attained results support this understanding: in the pipeline tuning phase, at least one DR-encompassing pipeline matched or surpassed the observed validation performance of the baseline pipelines for all classifiers. Albeit the p-values of the Friedman's test were below significance level, *post-hoc* testing demonstrated that this affected only a minority of pairwise comparisons. In fact, despite considering a representative set of techniques and p-value adjustment methods, the results did not support the rejection of the null hypothesis for the majority of pipeline comparisons.

In terms of the independent testing phase, a visual inspection of the produced projections attested that all projectors but UMAP-based ones seemed to converge on how to represent the small molecules in the projected space based on the fifty-dimensional data set originally produced. This happened for both two- and three-dimensional projections. Considering how projectors with different methodologies still converged on how to depict the fifty-dimensional data set, this augments the confidence regarding the fidelity of

observed patterns in the produced representations.

Although the performance of DR-encompassing pipelines diminished a bit in the independent testing phase, the ones for KNN and MLP still managed to surpass their baseline counterparts. For the remaining pipelines, most of them attained a comparable performance, further sustaining the understanding that DR via feature extraction can be relied on for classification purposes.

Between all projection approaches, Ivis was observed to recurrently surpass other techniques irrespective of the employed classifier. This could be attributed to the non-linear transformations it applies alongside the use of label information in its learning routine. However, considering how long Ivis-based pipelines took to fit the data when compared to others and how those based on linear projectors (e.g., PCA, TSVD, and PCS) still managed to perform competitively, there seems to be a trade-off between fitting time and prediction performance that needs to be evaluated by practitioners when deciding on which estimators to use to model a specific problem.

It is noteworthy that, since the applied BO procedure performed the same number of evaluations for all pipelines, those with a bigger number of HPs to tweak were at a disadvantage, as they had to optimize more HPs than pipelines with low-dimensional HP spaces with the same number of HP evaluations. In spite of that, models with a big number of HPs, such as those employing Ivis or UMAP, still managed to surpass baseline models on some occasions.

It is paramount to highlight that DR via feature extraction is not a one-size-fits-all solution: the generated projections, albeit representative of the entire original feature set, are composed of features that cannot be easily interpreted, precluding, for instance, the understanding of which features contributed the most to the produced model. In other words: pipelines that apply feature extraction are less interpretable. Furthermore, their addition to the pipeline brings an overhead that needs to be considered case-by-case. Ultimately, what needs to be internalized is that feature extraction is another tool in the practitioner's toolbox that can be relied upon just as much as feature selection in drug formulation and development.

In conclusion, this work successfully demonstrated how, for a particular context of biosciences, DR via feature extraction enables the application of algorithms that do not scale for high dimensionalities (e.g., scatter plots) and transforms the space in a way that potentially benefits ML tasks (e.g., classification). However, the decision of making use of DR, regardless of being via feature selection or extraction, should be a conscientious one that considers the idiosyncrasies of the problem domain and data at hand.

5.2 Future work

Throughout the development of this work, some aspects that can be further explored were acknowledged. To begin with, considering the wide spectrum of available molecular descriptors, a data set with a broader set of molecular descriptors could be procured. Molecular fingerprints could also be considered. Highly multivariate data sets, due to the curse of dimensionality, are where feature extraction techniques for DR can make their impact more perceptible. Considering how commonplace high-dimensional data sets are in many disciplines of biosciences, any improvements brought about by DR via feature extraction can have substantial impact in a multitude of fields.

In terms of the performed HPO procedure, one particularity is that it was governed by a fixed number of iterations that was equal regardless of the pipeline. This had to be made due to limitations in the BO implementation employed. Ideally, an adaptive stopping criteria—perhaps conditioned to the improvement rate of the best solution in the last iterations—could be adopted. Alternatively, a different optimization methodology could be employed, such as meta-heuristics like genetic algorithms and particle swarm optimizers. These changes should ensure that the optimization process only stops when no better solution is found instead of either abruptly stopping the optimization or letting it run beyond the necessary number of iterations. Any changes in this part, however, should be carefully considered, as any increase in terms of computational cost might further protract the overall runtime, thus diminishing the practical appeal of the proposed methodology.

Another improvement opportunity lies on the selection of DR techniques that comprised this work, which only considered feature extraction routines. It would be desirable to also cover a representative selection of popular and state-of-the-art feature selection algorithms to gain further insights on their respective impacts and compare feature selection against feature extraction. The set of classifiers used could also be enlarged to also cover novel approaches (e.g., Extreme Learning Machines).

Furthermore, it is important to note that all considered DR techniques but PCS support projections with more than three dimensions. Although projecting for higher dimensionalities eliminate the attractiveness of visualizing a representation of the entire set via conventional plotting techniques, representations with more dimensions could potentially retain more information or better represent the data, ultimately leading to better-performing classification models. In fact, considering how no DR needs to be performed for feature extraction to occur, one could employ projectors to merely attempt to improve the feature space while maintaining the original dimensionality.

Regarding the obtained results, one may safely conclude that the produced MLP-based pipelines are suffering performance issues. It is pertinent to investigate what is causing this behavior and improve the methodology to mitigate this issue and raise the

average performance of pipelines that employ MLP.

Lastly, the devised pipeline is applicable to a variety of other problems with minimal adaptation. As such, it would be interesting to see the results of applying it in related contexts, such as predicting brain barrier penetration of small molecules. A general-purpose web application could also be devised to allow practitioners of multiple disciplines to take advantage of this pipeline and methodology to solve problems in various science realms without the need to manipulate programming scripts and environments nor having to locally install required binaries.

UNDERGRADUATE CONTRIBUTIONS

This chapter enumerates the academic publications made in scientific journals and conferences throughout the completion of the bachelor's degree in computer science.

Scientific journal publications

- FLEXA, Caio; GOMES, Walisson; MOREIRA, Igor; ALVES, Ronnie; SALES, Claudomiro. Polygonal Coordinate System: Visualizing high-dimensional data using geometric DR, and a deterministic version of t-SNE. **Expert Systems with Applications**, v. 175, p. 114741, 2021. ISSN 0957-4174. DOI: https://doi.org/10.1016/j.eswa.2021.114741. Available from: <https://www.sciencedirect.com/science/article/pii/S0957417421001822>

Conference proceeding publications

- FLEXA, Caio; GOMES, Walisson; MOREIRA, Igor; SANTOS, Reginaldo; SALES, Claudomiro; SILVA, Moisés. Improving a Genetic Clustering Approach with a CVI-Based Objective Function. In_____. **Intelligent Systems**. Cham: Springer International Publishing, 2021. P. 202–217. ISBN 978-3-030-91702-9

- VASCONCELOS, Matheus; FLEXA, Caio; MOREIRA, Igor; SANTOS, Reginaldo; SALES, Claudomiro. Improving Particle Swarm Optimization with Self-adaptive Parameters, Rotational Invariance, and Diversity Control. In_____. **Intelligent Systems**. Cham: Springer International Publishing, 2021. P. 218–233. ISBN 978-3-030-91702-9

- BARRETO, Adriano; MOREIRA, Igor; FLEXA, Caio; CARDOSO, Eduardo; SALES, Claudomiro. An Online Pyramidal Embedding Technique for High Dimensional Big Data Visualization. In_____. **Intelligent Systems**. Cham: Springer International Publishing, 2020. P. 291–306. ISBN 978-3-030-61380-8

REFERENCES

AHMED, Ejaz; YAQOOB, Ibrar; HASHEM, Ibrahim Abaker Targio; KHAN, Imran; AHMED, Abdelmuttlib Ibrahim Abdalla; IMRAN, Muhammad; VASILAKOS, Athanasios V. The role of big data analytics in Internet of Things. **Computer Networks**, v. 129, p. 459–471, 2017. Special Issue on 5G Wireless Networks for IoT and Body Sensors. ISSN 1389-1286. Cit. on p. 22.

AHMED, Imran; AHMAD, Misbah; JEON, Gwanggil; PICCIALLI, Francesco. A Framework for Pandemic Prediction Using Big Data Analytics. en. **Big Data Research**, v. 25, p. 100190, July 2021. ISSN 22145796. DOI: 10.1016/j.bdr.2021.100190. Available from: <https://linkinghub.elsevier.com/retrieve/pii/S2214579621000071>. Visited on: 4 Apr. 2021. Cit. on p. 17.

ALNUAIMI, Noura; MASUD, Mohammad Mehedy; SERHANI, Mohamed Adel; ZAKI, Nazar. Streaming feature selection algorithms for big data: A survey. **Applied Computing and Informatics**, 2019. Cit. on p. 17.

ALSWAITTI, Mohammed; ALBUGHDADI, Mohanad; ISA, Nor Ashidi Mat. Density-based particle swarm optimization algorithm for data clustering. **Expert Systems with Applications**, v. 91, p. 170–186, 2018. ISSN 0957-4174. Cit. on p. 44.

AMRHEIN, Valentin; GREENLAND, Sander; MCSHANE, Blake. Scientists rise up against statistical significance. **Nature**, v. 567, p. 305–307, 21 Mar. 2019. Available from: <https://www.nature.com/articles/d41586-019-00857-9>. Cit. on p. 44.

ANG, Kenneth Li-Minn; GE, Feng Lu; SENG, Kah Phooi. Big Educational Data & Analytics: Survey, Architecture and Challenges. en. **IEEE Access**, v. 8, p. 116392–116414, 2020. ISSN 2169-3536. DOI: 10.1109/ACCESS.2020.2994561. Available from: <https://ieeexplore.ieee.org/document/9093868/>. Visited on: 4 Apr. 2021. Cit. on p. 22.

ANOWAR, Farzana; SADAOUI, Samira; SELIM, Bassant. Conceptual and empirical comparison of dimensionality reduction algorithms (PCA, KPCA, LDA, MDS, SVD, LLE, ISOMAP, LE, ICA, t-SNE). **Computer Science Review**, v. 40, p. 100378, May 2021. ISSN 15740137. DOI: 10.1016/j.cosrev.2021.100378. Available from: <https://linkinghub.elsevier.com/retrieve/pii/S1574013721000186>. Visited on: 24 Jan. 2022. Cit. on pp. 34, 36.

ASLAN, Muhammet Fatih; SABANCI, Kadir; DURDU, Akif; UNLERSEN, Muhammed Fahri. COVID-19 diagnosis using state-of-the-art CNN architecture features and Bayesian Optimization. **Computers in Biology and Medicine**, v. 142, p. 105244, Mar. 2022. ISSN 00104825. DOI: 10.1016/j.compbiomed.2022.105244. Available from: <https://linkinghub.elsevier.com/retrieve/pii/S0010482522000361>. Visited on: 13 Feb. 2022. Cit. on p. 43.

AYESHA, Shaeela; HANIF, Muhammad Kashif; TALIB, Ramzan. Overview and comparative study of dimensionality reduction techniques for high dimensional data. **Information Fusion**, v. 59, p. 44–58, July 2020. ISSN 15662535. DOI: 10.1016/j.inffus.2020.01.005. Available from: <https://linkinghub.elsevier.com/retrieve/pii/S156625351930377X>. Visited on: 24 Jan. 2022. Cit. on pp. 34, 35, 37.

BANNIGAN, Pauric; ALDEGHI, Matteo; BAO, Zeqing; HÄSE, Florian; ASPURU-GUZIK, Alán; ALLEN, Christine. Machine learning directed drug formulation development. **Advanced Drug Delivery Reviews**, v. 175, p. 113806, Aug. 2021. ISSN 0169409X. DOI: 10.1016/j.addr.2021.05.016. Available from: <https://linkinghub.elsevier.com/retrieve/pii/S0169409X21001800>. Visited on: 22 Jan. 2022. Cit. on pp. 17–19.

BARRETO, Adriano; MOREIRA, Igor; FLEXA, Caio; CARDOSO, Eduardo; SALES, Claudomiro. An Online Pyramidal Embedding Technique for High Dimensional Big Data Visualization. In_____. **Intelligent Systems**. Cham: Springer International Publishing, 2020. P. 291–306. ISBN 978-3-030-61380-8. Cit. on pp. 37, 50, 84.

BECHT, Etienne; MCINNES, Leland; HEALY, John; DUTERTRE, Charles-Antoine; KWOK, Immanuel WH; NG, Lai Guan; GINHOUX, Florent; NEWELL, Evan W. Dimensionality reduction for visualizing single-cell data using UMAP. **Nature biotechnology**, Nature Publishing Group, v. 37, n. 1, p. 38–44, 2019. Cit. on pp. 17, 38.

BELKIN, Mikhail; NIYOGI, Partha. Laplacian Eigenmaps and Spectral Techniques for Embedding and Clustering. In: ADVANCES in Neural Information Processing Systems. [S.l.]: MIT Press, 2002. v. 14. Available from: <https://proceedings.neurips.cc/paper/2001/hash/f106b7f99d2cb30c3db1c3cc0fde9ccb-Abstract.html>. Visited on: 14 Feb. 2022. Cit. on p. 38.

_____. Laplacian Eigenmaps for Dimensionality Reduction and Data Representation. **Neural Computation**, v. 15, n. 6, p. 1373–1396, 1 June 2003. ISSN 0899-7667, 1530-888X. DOI: 10.1162/089976603321780317. Available from: <https://direct.mit.edu/neco/article/15/6/1373-1396/6730>. Visited on: 14 Feb. 2022. Cit. on p. 38.

BENET, L. Z.; AL., et. BDDCS, the Rule of 5 and drugability. **Advanced Drug Delivery Reviews journa**, v. 101, p. 89–98, 2016. DOI: 10.1016/j.addr.2016.05.007. Cit. on p. 32.

BERGSTRA, James; BARDENET, Rémi; BENGIO, Yoshua; KÉGL, Balázs. Algorithms for Hyper-Parameter Optimization. In: ADVANCES in Neural Information Processing Systems. [S.l.]: Curran Associates, Inc., 2011. v. 24. Available from: <https://proceedings.neurips.cc/paper/2011/hash/86e8f7ab32cfd12577bc2619bc635690-Abstract.html>. Visited on: 13 Feb. 2022. Cit. on p. 42.

BICKERTON, G. R.; AL., et. Quantifying the chemical beauty of drugs. **Nature Chemistry**, v. 4, 2012. DOI: 10.1038/NCHEM.1243. Cit. on p. 31.

BOSER, Bernhard E.; GUYON, Isabelle M.; VAPNIK, Vladimir N. A Training Algorithm for Optimal Margin Classifiers. In: PROCEEDINGS of the Fifth Annual Workshop on Computational Learning Theory. New York, NY, USA: Association for Computing Machinery, 1992. (COLT '92), p. 144–152. event-place: Pittsburgh, Pennsylvania, USA. ISBN 0-89791-497-X. DOI: 10.1145/130385.130401. Available from: <https://doi.org/10.1145/130385.130401>. Cit. on pp. 40, 41.

BREIMAN, Leo. Random Forests. **Machine Learning**, v. 45, n. 1, p. 5–32, 1 Oct. 2001. ISSN 1573-0565. DOI: 10.1023/A:1010933404324. Available from: <https://doi.org/10.1023/A:1010933404324>. Cit. on pp. 40, 41.

BRUNELLO, Andrea; JIMÉNEZ, Fernando; MARZANO, Enrico; MONTANARI, Angelo; SÁNCHEZ, Gracia; SCIAVICCO, Guido. Multiobjective evolutionary feature selection and fuzzy classification of contact centre data. **Expert Systems**, v. 36, Mar. 2019. Cit. on p. 34.

BURGES, Christopher JC. **Dimension reduction: A guided tour**. [S.l.]: Now Publishers Inc, 2010. Cit. on pp. 18, 33, 34.

CARRACEDO-REBOREDO, Paula; LIÑARES-BLANCO, Jose; RODRÍGUEZ-FERNÁNDEZ, Nereida; CEDRÓN, Francisco; NOVOA, Francisco J.; CARBALLAL, Adrian; MAOJO, Victor; PAZOS, Alejandro; FERNANDEZ-LOZANO, Carlos. A review on machine learning approaches and trends in drug discovery. **Computational and Structural Biotechnology Journal**, v. 19, p. 4538–4558, 2021. ISSN 20010370. DOI: 10.1016/j.csbj.2021.08.011. Available from: <https://linkinghub.elsevier.com/retrieve/pii/S2001037021003421>. Visited on: 22 Jan. 2022. Cit. on pp. 18–21, 23, 29–32.

CARRASCO-CORREA, Enrique Javier; RUIZ-ALLICA, Julia; RODRÍGUEZ-FERNÁNDEZ, Juan Francisco; MIRÓ, Manuel. Human artificial membranes in (bio)analytical science: Potential for in vitro prediction of intestinal absorption-A review. **TrAC Trends in Analytical Chemistry**, v. 145, p. 116446, Dec.

2021. ISSN 01659936. DOI: 10.1016/j.trac.2021.116446. Available from: <https://linkinghub.elsevier.com/retrieve/pii/S0165993621002697>. Visited on: 24 Jan. 2022. Cit. on p. 28.

CASTELLETTI, A.; LOTOV, A.V.; SONCINI-SESSA, R. Visualization-based multi-objective improvement of environmental decision-making using linearization of response surfaces. **Environmental Modelling & Software**, v. 25, n. 12, p. 1552–1564, 2010. ISSN 1364-8152. Cit. on p. 17.

CHAFFI, Babak Nasseh; TAFRESHI, Fakhteh Soltani. Nasseh method to visualize high-dimensional data. **Applied Soft Computing**, v. 84, 2019. Cit. on p. 24.

CHEN, C.L. Philip; ZHANG, Chun-Yang. Data-intensive applications, challenges, techniques and technologies: A survey on Big Data. **Information Sciences**, v. 275, p. 314–347, 2014. Cit. on pp. 17, 18, 22, 23.

CHENG, P.; LI, W.; OGUNBONA, P. Greedy Approximation of Kernel PCA by Minimizing the Mapping Error. In: DIGITAL Image Computing: Techniques and Applications, 2009. DICTA '09. [S.l.: s.n.], Dec. 2009. P. 303–308. Cit. on p. 37.

CONOVER, W.J. **Practical Nonparametric Statistics**. [S.l.]: Wiley, 1999. (Wiley Series in Probability and Statistics). ISBN 978-0-471-16068-7. Available from: <https://books.google.com.br/books?id=n%5C_39DwAAQBAJ>. Cit. on pp. 43, 44.

COOPER, Robert A. Making Decisions with Data: Understanding Hypothesis Testing & Statistical Significance. **The American Biology Teacher**, v. 81, n. 8, p. 535–542, 1 Oct. 2019. ISSN 0002-7685, 1938-4211. DOI: 10.1525/abt.2019.81.8.535. Available from: <https://online.ucpress.edu/abt/article/81/8/535/20710/Making-Decisions-with-Data-Understanding>. Visited on: 14 Feb. 2022. Cit. on p. 44.

CUNNINGHAM, John P; GHAHRAMANI, Zoubin. Linear dimensionality reduction: Survey, insights, and generalizations. **The Journal of Machine Learning Research**, JMLR. org, v. 16, n. 1, p. 2859–2900, 2015. Cit. on p. 35.

D'ALCONZO, Alessandro; DRAGO, Idilio; MORICHETTA, Andrea; MELLIA, Marco; CASAS, Pedro. A Survey on Big Data for Network Traffic Monitoring and Analysis. en. **IEEE Transactions on Network and Service Management**, v. 16, n. 3, p. 800–813, Sept. 2019. ISSN 1932-4537, 2373-7379. DOI: 10.1109/TNSM.2019.2933358. Available from: <https://ieeexplore.ieee.org/document/8789667/>. Visited on: 4 Apr. 2021. Cit. on pp. 21, 23.

DAHLGREN, David. **Biopharmaceutical aspects of intestinal drug absorption: Regional permeability and absorption-modifying excipients**. 2018. PhD thesis – UPPSALA UNIVERSITY. Cit. on p. 30.

DAHLGREN, David; LENNERNÄS, Hans. Intestinal Permeability and Drug Absorption: Predictive Experimental, Computational and In Vivo Approaches. **Pharmaceutics**, v. 11, n. 8, p. 411, 13 Aug. 2019. ISSN 1999-4923. DOI: 10.3390/pharmaceutics11080411. Available from: <https://www.mdpi.com/1999-4923/11/8/411>. Visited on: 25 Jan. 2022. Cit. on pp. 18, 20, 21.

DAINA, Antoine; ZOETE, Vincent. A BOILED-Egg To Predict Gastrointestinal Absorption and Brain Penetration of Small Molecules. **Chemmedchem**, v. 11, n. 11, p. 1117–1121, 6 June 2016. ISSN 1860-7179. DOI: 10.1002/cmdc.201600182. Available from: <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5089604/>. Visited on: 11 Feb. 2022. Cit. on pp. 24, 28.

DANIEL, W. W. **Applied nonparametric statistics**. Boston (Mass.), USA: PWS-KENT, 1990. Cit. on p. 44.

DANISHUDDIN; KUMAR, Vikas; FAHEEM, Mohammad; WOO LEE, Keun. A decade of machine learning-based predictive models for human pharmacokinetics: Advances and challenges. **Drug Discovery Today**, s1359644621004074, Sept. 2021. ISSN 13596446. DOI: 10.1016/j.drudis.2021.09.013. Available from: <https://linkinghub.elsevier.com/retrieve/pii/S1359644621004074>. Visited on: 22 Jan. 2022. Cit. on pp. 17–20, 22.

DEMŠAR, Janez. Statistical Comparisons of Classifiers over Multiple Data Sets. **The Journal of Machine Learning Research**, v. 7, p. 1–30, Jan. 2006. ISSN 1532-4435. Available from: <https://www.jmlr.org/papers/volume7/demsar06a/demsar06a.pdf>. Cit. on p. 44.

DIX, Alan. Human–computer interaction, foundations and new paradigms. **Journal of Visual Languages & Computing**, v. 42, p. 122–134, 2017. Cit. on p. 17.

DROZDOV, Ignat; FORBES, Daniel; SZUBERT, Benjamin; HALL, Mark; CARLIN, Chris; LOWE, David J. Supervised and unsupervised language modelling in Chest X-Ray radiological reports. Ed. by Ulas Bagci. **PLOS ONE**, v. 15, n. 3, e0229963, 10 Mar. 2020. ISSN 1932-6203. DOI: 10.1371/journal.pone.0229963. Available from: <https://dx.plos.org/10.1371/journal.pone.0229963>. Visited on: 25 Jan. 2022. Cit. on p. 38.

DRYDEN, Ian L.; HODGE, David J. Journeys in big data statistics. **Statistics & Probability Letters**, v. 136, p. 121–125, 2018. The role of Statistics in the era of big data. Cit. on p. 18.

ELBADAWI, Moe; GAISFORD, Simon; BASIT, Abdul W. Advanced machine-learning techniques in drug discovery. **Drug Discovery Today**, v. 26, n. 3, p. 769–777, Mar. 2021. ISSN 13596446. DOI: 10.1016/j.drudis.2020.12.003. Available from:

<https://linkinghub.elsevier.com/retrieve/pii/S1359644620305213>. Visited on: 22 Jan. 2022. Cit. on pp. 20, 23.

ENGEL, Daniel; HÜTTENBERGER, Lars; HAMANN, Bernd. A survey of dimension reduction methods for high-dimensional data analysis and visualization. In: SCHLOSS DAGSTUHL-LEIBNIZ-ZENTRUM FUER INFORMATIK. VISUALIZATION of Large and Unstructured Data Sets: Applications in Geospatial Planning, Modeling and Engineering-Proceedings of IRTG 1131 Workshop 2011. [S.l.: s.n.], 2012. Cit. on pp. 18, 22, 33, 34.

ESAKI, Tsuyoshi; OHASHI, Rikiya; WATANABE, Reiko; NATSUME-KITATANI, Yayoi; KAWASHIMA, Hitoshi; NAGAO, Chioko; KOMURA, Hiroshi; MIZUGUCHI, Kenji. Constructing an In Silico Three-Class Predictor of Human Intestinal Absorption With Caco-2 Permeability and Dried-DMSO Solubility. **Journal of Pharmaceutical Sciences**, v. 108, n. 11, p. 3630–3639, Nov. 2019. ISSN 00223549. DOI: 10.1016/j.xphs.2019.07.014. Available from: <https://linkinghub.elsevier.com/retrieve/pii/S0022354919304496>. Visited on: 23 Jan. 2022. Cit. on pp. 21, 25, 26, 40.

ESPADOTO, Mateus; MARTINS, Rafael M; KERREN, Andreas; HIRATA, Nina ST; TELEA, Alexandru Cristian. Towards a quantitative survey of dimension reduction techniques. **IEEE Transactions on Visualization and Computer Graphics**, IEEE, 2019. Cit. on pp. 18, 24, 33–35.

FARRÉ, Ricard; FIORANI, Marcello; ABDU RAHIMAN, Saeed; MATTEOLI, Gianluca. Intestinal Permeability, Inflammation and the Role of Nutrients. **Nutrients**, v. 12, n. 4, p. 1185, 23 Apr. 2020. ISSN 2072-6643. DOI: 10.3390/nu12041185. Available from: <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7231157/>. Visited on: 1 Feb. 2022. Cit. on p. 18.

FAYYAD, Usama. From Data Mining to Knowledge Discovery in Databases. en. **AI magazine**, v. 17, p. 18, 1996. DOI: https://doi.org/10.1609/aimag.v17i3.1230. Available from: <https://ojs.aaai.org/index.php/aimagazine/article/view/1230>. Cit. on pp. 17, 21, 22.

FEDI, Arianna; VITALE, Chiara; PONSCHIN, Giulia; AYEHUNIE, Seyoum; FATO, Marco; SCAGLIONE, Silvia. In vitro models replicating the human intestinal epithelium for absorption and metabolism studies: A systematic review. **Journal of Controlled Release**, v. 335, p. 247–268, July 2021. ISSN 01683659. DOI: 10.1016/j.jconrel.2021.05.028. Available from: <https://linkinghub.elsevier.com/retrieve/pii/S0168365921002522>. Visited on: 24 Jan. 2022. Cit. on pp. 20, 21, 29, 30.

FENG, Mingchen; ZHENG, Jiangbin; REN, Jinchang; HUSSAIN, Amir; LI, Xiuxiu; XI, Yue; LIU, Qiaoyuan. Big Data Analytics and Mining for Effective Visualization and Trends Forecasting of Crime Data. en. **IEEE Access**, v. 7, p. 106111–106123, 2019. ISSN 2169-3536. DOI: 10.1109/ACCESS.2019.2930410. Available from: <https://ieeexplore.ieee.org/document/8768367/>. Visited on: 4 Apr. 2021. Cit. on p. 23.

FERNÁNDEZ, Alicia; GÓMEZ, Álvaro; LECUMBERRY, Federico; PARDO, Álvaro; RAMÍREZ, Ignacio. Pattern Recognition in Latin America in the "Big Data" Era. **Pattern Recognition**, v. 48, n. 4, p. 1185–1196, 2015. Cit. on pp. 22, 44.

FERNANDEZ-DELGADO, Manuel; CERNADAS, Eva; BARRO, Senen; AMORIM, Dinani. Do we Need Hundreds of Classifiers to Solve Real World Classification Problems?, p. 49, 2014. Available from: <https://www.jmlr.org/papers/volume15/delgado14a/delgado14a.pdf>. Cit. on pp. 33, 40, 41.

FERNÁNDEZ-TORRAS, Adrià; COMAJUNCOSA-CREUS, Arnau; DURAN-FRIGOLA, Miquel; ALOY, Patrick. Connecting chemistry and biology through molecular descriptors. **Current Opinion in Chemical Biology**, v. 66, p. 102090, Feb. 2022. ISSN 13675931. DOI: 10.1016/j.cbpa.2021.09.001. Available from: <https://linkinghub.elsevier.com/retrieve/pii/S1367593121001204>. Visited on: 19 Feb. 2022. Cit. on p. 30.

FIX, Evelyn; HODGES, JL. **Nonparametric Discrimination: Consistency Properties**. Randolph Field, Texas, Feb. 1951. Available from: <https://apps.dtic.mil/dtic/tr/fulltext/u2/a800276.pdf>. Cit. on p. 40.

FLEXA, C.; GOMES, W.; VIADEMONTE, S.; SALES, C.; ALVES, R. A geometry-based approach to visualize high-dimensional data. In: 2019 8th Brazilian Conference on Intelligent Systems (BRACIS). [S.l.: s.n.], Oct. 2019. Cit. on pp. 35, 36.

FLEXA, Caio; GOMES, Walisson; MOREIRA, Igor; ALVES, Ronnie; SALES, Claudomiro. Polygonal Coordinate System: Visualizing high-dimensional data using geometric DR, and a deterministic version of t-SNE. **Expert Systems with Applications**, v. 175, p. 114741, 2021. ISSN 0957-4174. DOI: https://doi.org/10.1016/j.eswa.2021.114741. Available from: <https://www.sciencedirect.com/science/article/pii/S0957417421001822>. Cit. on pp. 24, 37, 38, 44, 84.

FLEXA, Caio; GOMES, Walisson; MOREIRA, Igor; SANTOS, Reginaldo; SALES, Claudomiro; SILVA, Moisés. Improving a Genetic Clustering Approach with a CVI-Based Objective Function. In_____. **Intelligent Systems**. Cham: Springer International Publishing, 2021. P. 202–217. ISBN 978-3-030-91702-9. Cit. on p. 84.

FLEXA, Caio; SANTOS, Reginaldo; GOMES, Walisson; SALES, Claudomiro; COSTA, João C.W.A. Mutual equidistant-scattering criterion: A new index for crisp clustering. **Expert Systems with Applications**, v. 128, p. 225–245, 2019. Cit. on p. 34.

FU, Yuankun; SONG, Fengguang; ZHU, Luoding. Modeling and Implementation of an Asynchronous Approach to Integrating HPC and Big Data Analysis. **Procedia Computer Science**, v. 80, p. 52–62, 2016. Cit. on p. 17.

GARCÍA, Salvador; FERNÁNDEZ, Alberto; LUENGO, Julián; HERRERA, Francisco. Advanced nonparametric tests for multiple comparisons in the design of experiments in computational intelligence and data mining: Experimental analysis of power. **Information Sciences**, v. 180, n. 10, p. 2044–2064, 2010. Special Issue on Intelligent Distributed Information Systems. Cit. on p. 44.

GARDINER, Adrian; AASHEIM, Cheryl; RUTNER, Paige; WILLIAMS, Susan. Skill Requirements in Big Data: A Content Analysis of Job Advertisements. **Journal of Computer Information Systems**, Taylor & Francis, v. 58, n. 4, p. 374–384, 2018. Cit. on p. 22.

GATARIĆ, Biljana; PAROJČIĆ, Jelena. An Investigation into the Factors Governing Drug Absorption and Food Effect Prediction Based on Data Mining Methodology. **The AAPS Journal**, v. 22, n. 1, p. 11, Jan. 2020. ISSN 1550-7416. DOI: 10.1208/s12248-019-0394-y. Available from: <http://link.springer.com/10.1208/s12248-019-0394-y>. Visited on: 8 Feb. 2022. Cit. on pp. 25, 26.

GAVINS, Francesca K.H.; FU, Zihao; ELBADAWI, Moe; BASIT, Abdul W.; RODRIGUES, Miguel R.D.; ORLU, Mine. Machine learning predicts the effect of food on orally administered medicines. **International Journal of Pharmaceutics**, v. 611, p. 121329, Jan. 2022. ISSN 03785173. DOI: 10.1016/j.ijpharm.2021.121329. Available from: <https://linkinghub.elsevier.com/retrieve/pii/S0378517321011352>. Visited on: 22 Jan. 2022. Cit. on pp. 25, 26, 28.

GIULIANI, Alessandro. The application of principal component analysis to drug discovery and biomedical data. **Drug Discovery Today**, v. 22, n. 7, p. 1069–1076, July 2017. ISSN 13596446. DOI: 10.1016/j.drudis.2017.01.005. Available from: <https://linkinghub.elsevier.com/retrieve/pii/S1359644617300181>. Visited on: 25 Jan. 2022. Cit. on p. 36.

GODDARD, Erin; KLEIN, Colin; SOLOMON, Samuel G.; HOGENDOORN, Hinze; CARLSON, Thomas A. Interpreting the dimensions of neural feature representations revealed by dimensionality reduction. **NeuroImage**, v. 180, p. 41–67, 2018. Cit. on p. 34.

GOPI, E. S.; PALANISAMY, P. Neural network based class-conditional probability density function using kernel trick for supervised classifier. **Neurocomputing**, v. 154, p. 225–229, 2015. Cit. on p. 37.

GRIFFITHS, Peter; NEEDLEMAN, Jack. Statistical significance testing and p-values: Defending the indefensible? A discussion paper and position statement. **International Journal of Nursing Studies**, v. 99, p. 103384, Nov. 2019. ISSN 00207489. DOI: 10.1016/j.ijnurstu.2019.07.001. Available from: <https://linkinghub.elsevier.com/retrieve/pii/S0020748919301774>. Visited on: 14 Feb. 2022. Cit. on p. 44.

GUPTA, Deepak; RANI, Rinkle. Improving malware detection using big data and ensemble learning. en. **Computers & Electrical Engineering**, v. 86, p. 106729, Sept. 2020. ISSN 00457906. DOI: 10.1016/j.compeleceng.2020.106729. Available from: <https://linkinghub.elsevier.com/retrieve/pii/S004579062030584X>. Visited on: 4 Apr. 2021. Cit. on p. 23.

HABEEB, Riyaz Ahamed Ariyaluran; NASARUDDIN, Fariza; GANI, Abdullah; HASHEM, Ibrahim Abaker Targio; AHMED, Ejaz; IMRAN, Muhammad. Real-time big data processing for anomaly detection: A Survey. **International Journal of Information Management**, v. 45, p. 289–307, 2019. Cit. on pp. 17, 22.

HALKO, N; MARTINSSON, P G; TROPP, J A. FINDING STRUCTURE WITH RANDOMNESS: STOCHASTIC ALGORITHMS FOR CONSTRUCTING APPROXIMATE MATRIX DECOMPOSITIONS. **Applied & Computational Mathematics**, p. 82, Sept. 2009. Cit. on pp. 35, 36.

HAN, Yongming; SONG, Guangliang; LIU, Fenfen; GENG, Zhiqiang; MA, Bo; XU, Wei. Fault monitoring using novel adaptive kernel principal component analysis integrating grey relational analysis. **Process Safety and Environmental Protection**, v. 157, p. 397–410, Jan. 2022. ISSN 09575820. DOI: 10.1016/j.psep.2021.11.029. Available from: <https://linkinghub.elsevier.com/retrieve/pii/S0957582021006352>. Visited on: 25 Jan. 2022. Cit. on p. 37.

HANSON, Richard J. A Numerical Method for Solving Fredholm Integral Equations of the First Kind Using Singular Values. **SIAM Journal on Numerical Analysis**, v. 8, n. 3, p. 616–622, 1971. Publisher: Society for Industrial and Applied Mathematics. ISSN 0036-1429. Available from: <https://www.jstor.org/stable/2949679>. Visited on: 12 Feb. 2022. Cit. on p. 35.

HARRIS, Charles R.; MILLMAN, K. Jarrod; WALT, Stéfan J. van der; GOMMERS, Ralf; VIRTANEN, Pauli; COURNAPEAU, David; WIESER, Eric; TAYLOR, Julian; BERG, Sebastian; SMITH, Nathaniel J.; KERN, Robert; PICUS, Matti; HOYER, Stephan; KERKWIJK, Marten H. van; BRETT, Matthew; HALDANE, Allan; RÍO, Jaime Fernández del; WIEBE, Mark; PETERSON, Pearu;

GÉRARD-MARCHANT, Pierre; SHEPPARD, Kevin; REDDY, Tyler; WECKESSER, Warren; ABBASI, Hameer; GOHLKE, Christoph; OLIPHANT, Travis E. Array programming with NumPy. **Nature**, Springer Science and Business Media LLC, v. 585, n. 7825, p. 357–362, Sept. 2020. DOI: 10.1038/s41586-020-2649-2. Available from: <https://doi.org/10.1038/s41586-020-2649-2>. Cit. on p. 49.

HAYKIN, S. **Neural Networks and Learning Machines**. [S.l.]: Pearson Prentice Hall New Jersey, 2008. Cit. on p. 40.

HE, Sheng; LEANSE, Leon G.; FENG, Yanfang. Artificial intelligence and machine learning assisted drug delivery for effective treatment of infectious diseases. **Advanced Drug Delivery Reviews**, v. 178, p. 113922, Nov. 2021. ISSN 0169409X. DOI: 10.1016/j.addr.2021.113922. Available from: <https://linkinghub.elsevier.com/retrieve/pii/S0169409X2100315X>. Visited on: 22 Jan. 2022. Cit. on pp. 20, 23.

HOLIDAY, Alexander; KOOSHKBAGHI, Mahdi; BELLO-RIVAS, Juan M.; GEAR, C. William; ZAGARIS, Antonios; KEVREKIDIS, Ioannis G. Manifold learning for parameter reduction. **Journal of Computational Physics**, v. 392, p. 419–431, 2019. Cit. on p. 24.

HOUARI, Rima; BOUNCEUR, Ahcène; KECHADI, M-Tahar; TARI, A-Kamel; EULER, Reinhardt. Dimensionality reduction in data mining: A Copula approach. **Expert Systems with Applications**, Elsevier, v. 64, p. 247–260, 2016. Cit. on p. 24.

_____._____. **Expert Systems with Applications**, v. 64, p. 247–260, 2016. Cit. on pp. 34, 35.

HOULE, Michael E. Local Intrinsic Dimensionality II: Multivariate Analysis and Distributional Support. In: SIMILARITY Search and Applications. Cham: Springer International Publishing, 2017. P. 80–95. Cit. on p. 34.

HOULE, Michael E.; SCHUBERT, Erich; ZIMEK, Arthur. On the Correlation Between Local Intrinsic Dimensionality and Outlierness. In: SIMILARITY Search and Applications. Cham: Springer International Publishing, 2018. P. 177–191. Cit. on p. 34.

HUNTER, J. D. Matplotlib: A 2D graphics environment. **Computing in Science & Engineering**, IEEE COMPUTER SOC, v. 9, n. 3, p. 90–95, 2007. DOI: 10.1109/MCSE.2007.55. Cit. on p. 46.

HUSSAIN, Malik Zawwar; HUSSAIN, Maria. Visualization of data preserving monotonicity. **Applied Mathematics and Computation**, v. 190, n. 2, p. 1353–1364, 2007. ISSN 0096-3003. Cit. on pp. 17, 22.

ISLAM, Md Tauhidul; XING, Lei. A data-driven dimensionality-reduction algorithm for the exploration of patterns in biomedical data. **Nature Biomedical Engineering**, v. 5, n. 6, p. 624–635, June 2021. ISSN 2157-846X. DOI: 10.1038/s41551-020-00635-3. Available from: <http://www.nature.com/articles/s41551-020-00635-3>. Visited on: 24 Jan. 2022. Cit. on p. 38.

JAMES, Gareth; WITTEN, Daniela; HASTIE, Trevor; TIBSHIRANI, Robert. **An introduction to statistical learning**. [S.l.]: Springer, 2013. v. 112. Cit. on p. 41.

JIANG, Rong; LU, Rongxing; CHOO, Kim-Kwang Raymond. Achieving high performance and privacy-preserving query over encrypted multidimensional big metering data. **Future Generation Computer Systems**, v. 78, p. 392–401, 2018. Cit. on p. 18.

JIN, Xiaolong; WAH, Benjamin W.; CHENG, Xueqi; WANG, Yuanzhuo. Significance and Challenges of Big Data Research. **Big Data Research**, v. 2, n. 2, p. 59–64, 2015. Visions on Big Data. ISSN 2214-5796. Cit. on pp. 17, 22.

JOLLIFFE, I. T. **Principal Component Analysis**. New York: Springer-Verlag, 2002. (Springer Series in Statistics). Cit. on p. 36.

JONES, L.V. **The Collected Works of John W. Tukey: Philosophy and Principles of Data Analysis 1965-1986**. [S.l.]: Taylor & Francis, 1987. ISBN 978-0-534-05101-3. Available from: <https://books.google.com.br/books?id=C1guHWTlVVoC>. Cit. on p. 43.

KAMIYA, Yusuke; OMURA, Asuka; HAYASAKA, Riku; SAITO, Rie; SANO, Izumi; HANDA, Kentaro; OHORI, Junya; KITAJIMA, Masato; SHONO, Fumiaki; FUNATSU, Kimito; YAMAZAKI, Hiroshi. Prediction of permeability across intestinal cell monolayers for 219 disparate chemicals using in vitro experimental coefficients in a pH gradient system and in silico analyses by trivariate linear regressions and machine learning. **Biochemical Pharmacology**, v. 192, p. 114749, Oct. 2021. ISSN 00062952. DOI: 10.1016/j.bcp.2021.114749. Available from: <https://linkinghub.elsevier.com/retrieve/pii/S0006295221003658>. Visited on: 23 Jan. 2022. Cit. on p. 25.

KANG, Zhao; PENG, Chong; CHENG, Qiang. Kernel-driven similarity learning. **Neurocomputing**, v. 267, p. 210–219, 2017. Cit. on p. 37.

KIM, Kyoungok. An improved semi-supervised dimensionality reduction using feature weighting: Application to sentiment analysis. **Expert Systems with Applications**, v. 109, p. 49–65, 2018. Cit. on pp. 34, 35.

KIM, Kyoungok; LEE, Jaewook. Sentiment visualization and classification via semi-supervised nonlinear dimensionality reduction. **Pattern Recognition**, v. 47, n. 2, p. 758–768, 2014. ISSN 0031-3203. Cit. on p. 18.

KIM, Kyoungok; LEE, Jaewook. Sequential manifold learning for efficient churn prediction. **Expert Systems with Applications**, v. 39, n. 18, p. 13328–13337, 2012. Cit. on p. 35.

KOBAK, Dmitry; LINDERMAN, George C. Initialization is critical for preserving global data structure in both t-SNE and UMAP. **Nature Biotechnology**, v. 39, n. 2, p. 156–157, Feb. 2021. ISSN 1087-0156, 1546-1696. DOI: 10.1038/s41587-020-00809-z. Available from: <http://www.nature.com/articles/s41587-020-00809-z>. Visited on: 14 Feb. 2022. Cit. on p. 38.

_____. **UMAP does not preserve global structure any better than t-SNE when using the same initialization**. [S.l.], 19 Dec. 2019. DOI: 10.1101/2019.12.19.877522. Available from: <http://biorxiv.org/lookup/doi/10.1101/2019.12.19.877522>. Visited on: 14 Feb. 2022. Cit. on p. 38.

KUMAR, Rajnish; SHARMA, Anju; SIDDIQUI, Mohammed Haris; TIWARI, Rajesh Kumar. Prediction of Human Intestinal Absorption of Compounds Using Artificial Intelligence Techniques. **Current Drug Discovery Technologies**, v. 14, n. 4, 27 Oct. 2017. ISSN 15701638. DOI: 10.2174/1570163814666170404160911. Available from: <http://www.eurekaselect.com/151330/article>. Visited on: 23 Jan. 2022. Cit. on pp. 18, 19, 21, 25, 30, 40.

L'HEUREUX, Alexandra; GROLINGER, Katarina; ELYAMANY, Hany F.; CAPRETZ, Miriam A. M. Machine Learning With Big Data: Challenges and Approaches. en. **IEEE Access**, v. 5, p. 7776–7797, 2017. ISSN 2169-3536. DOI: 10.1109/ACCESS.2017.2696365. Available from: <https://ieeexplore.ieee.org/document/7906512/>. Visited on: 4 Apr. 2021. Cit. on p. 23.

LANDRUM, Greg. **RDKit: Open-source cheminformatics**. [S.l.: s.n.], 2020. DOI: 10.5281/zenodo.3732262. Available from: <http://www.rdkit.org>. Cit. on p. 48.

LANEY, Doug. 3D data management: Controlling data volume, velocity and variety. **META group research note**, v. 6, n. 70, p. 1, 2001. Cit. on pp. 18, 22.

LEE, Ming-Han; TA, Giang Huong; WENG, Ching-Feng; LEONG, Max K. In Silico Prediction of Intestinal Permeability by Hierarchical Support Vector Regression. **International Journal of Molecular Sciences**, v. 21, n. 10, p. 3582, 19 May 2020. ISSN 1422-0067. DOI: 10.3390/ijms21103582. Available from: <https://www.mdpi.com/1422-0067/21/10/3582>. Visited on: 23 Jan. 2022. Cit. on pp. 21, 25.

LI, Bo; LI, Yan-Rui; ZHANG, Xiao-Long. A survey on Laplacian eigenmaps based manifold learning methods. **Neurocomputing**, v. 335, p. 336–351, 2019. Cit. on pp. 24, 34, 35.

LI, Fangyu; XIE, Rui; WANG, Zengyan; GUO, Lulu; YE, Jin; MA, Ping; SONG, Wenzhan. Online Distributed IoT Security Monitoring With Multidimensional Streaming Big Data. en. **IEEE Internet of Things Journal**, v. 7, n. 5, p. 4387–4394, May 2020. ISSN 2327-4662, 2372-2541. DOI: 10.1109/JIOT.2019.2962788. Available from: <https://ieeexplore.ieee.org/document/8944307/>. Visited on: 4 Apr. 2021. Cit. on p. 22.

LI, Yinhui; XIA, Jingbo; ZHANG, Silan; YAN, Jiakai; AI, Xiaochuan; DAI, Kuobin. An efficient intrusion detection system based on support vector machines and gradually feature removal method. **Expert Systems with Applications**, v. 39, n. 1, p. 424–430, 2012. Cit. on p. 34.

LINDERMAN, George C; RACHH, Manas; HOSKINS, Jeremy G; STEINERBERGER, Stefan; KLUGER, Yuval. Fast interpolation-based t-SNE for improved visualization of single-cell RNA-seq data. **Nature methods**, Nature Publishing Group, v. 16, n. 3, p. 243–245, 2019. Cit. on p. 17.

LING, Yongsheng; YUE, Qi; CHAI, Chaojun; SHAN, Qing; HEI, Daqian; JIA, Wenbao. Nuclear accident source term estimation using Kernel Principal Component Analysis, Particle Swarm Optimization, and Backpropagation Neural Networks. **Annals of Nuclear Energy**, v. 136, p. 107031, Feb. 2020. ISSN 03064549. DOI: 10.1016/j.anucene.2019.107031. Available from: <https://linkinghub.elsevier.com/retrieve/pii/S030645491930533X>. Visited on: 25 Jan. 2022. Cit. on p. 37.

LIPINSKI, Christopher A; LOMBARDO, Franco; DOMINY, Beryl W; FEENEY, Paul J. Experimental and computational approaches to estimate solubility and permeability in drug discovery and development settings. **Advanced drug delivery reviews**, Elsevier, v. 23, n. 1-3, p. 3–25, 1997. Cit. on p. 31.

LIU, Yanan; LI, Na; ZHU, Xiao; QI, Yi. How wide is the application of genetic big data in biomedicine. en. **Biomedicine & Pharmacotherapy**, v. 133, p. 111074, Jan. 2021. ISSN 07533322. DOI: 10.1016/j.biopha.2020.111074. Available from: <https://linkinghub.elsevier.com/retrieve/pii/S0753332220312671>. Visited on: 4 Apr. 2021. Cit. on p. 17.

LIU, Z.; WANG, J.; LIU, G.; PU, J. Sparse Low-Rank Preserving Projection for Dimensionality Reduction. **IEEE Access**, v. 7, p. 22941–22951, 2019. Cit. on p. 34.

LO, Yu-Chen; RENSI, Stefano E.; TORNG, Wen; ALTMAN, Russ B. Machine learning in chemoinformatics and drug discovery. **Drug Discovery Today**, v. 23, n. 8, p. 1538–1546, Aug. 2018. ISSN 13596446. DOI: 10.1016/j.drudis.2018.05.010. Available from: <https://linkinghub.elsevier.com/retrieve/pii/S1359644617304695>. Visited on: 8 Feb. 2022. Cit. on pp. 20, 23, 31, 32, 35.

LOEY, Mohamed; EL-SAPPAGH, Shaker; MIRJALILI, Seyedali. Bayesian-based optimized deep learning model to detect COVID-19 patients using chest X-ray image data. **Computers in Biology and Medicine**, v. 142, p. 105213, Mar. 2022. ISSN 00104825. DOI: 10.1016/j.compbiomed.2022.105213. Available from: <https://linkinghub.elsevier.com/retrieve/pii/S0010482522000051>. Visited on: 13 Feb. 2022. Cit. on p. 43.

LOVELL, David P. Null hypothesis significance testing and effect sizes: can we 'effect' everything ... or ... anything? **Current Opinion in Pharmacology**, v. 51, p. 68–77, Apr. 2020. ISSN 14714892. DOI: 10.1016/j.coph.2019.12.001. Available from: <https://linkinghub.elsevier.com/retrieve/pii/S1471489219301171>. Visited on: 14 Feb. 2022. Cit. on p. 44.

LOVERING, F.; BIKKER, J.; HUMBLET, C. Escape from Flatland: Increasing Saturation as an Approach to Improving Clinical Success. **J. Med. Chem.**, v. 52, p. 6752–6756, 2009. DOI: 10.1021/jm901241e. Cit. on p. 32.

LU, Ruqian; JIN, Xiaolong; ZHANG, Songmao; QIU, Meikang; WU, Xindong. A Study on Big Knowledge and Its Engineering Issues. en. **IEEE Transactions on Knowledge and Data Engineering**, v. 31, n. 9, p. 1630–1644, Sept. 2019. ISSN 1041-4347, 1558-2191, 2326-3865. DOI: 10.1109/TKDE.2018.2866863. Available from: <https://ieeexplore.ieee.org/document/8444740/>. Visited on: 4 Apr. 2021. Cit. on pp. 19, 21.

LV, Zhihan; SONG, Houbing; BASANTA-VAL, Pablo; STEED, Anthony; JO, Minho. Next-Generation Big Data Analytics: State of the Art, Challenges, and Future Research Topics. en. **IEEE Transactions on Industrial Informatics**, v. 13, n. 4, p. 1891–1899, Aug. 2017. ISSN 1551-3203, 1941-0050. DOI: 10.1109/TII.2017.2650204. Available from: <https://ieeexplore.ieee.org/document/7866003/>. Visited on: 4 Apr. 2021. Cit. on p. 23.

MAATEN, Laurens van der. Accelerating t-SNE using Tree-Based Algorithms. **Journal of Machine Learning Research**, v. 15, p. 3221–3245, 2014. Cit. on pp. 23, 37.

MAATEN, Laurens van der; HINTON, Geoffrey. Visualizing Data using t-SNE. **Journal of Machine Learning Research**, v. 9, p. 2579–2605, 2008. Cit. on pp. 35, 37, 38.

MAATEN, Laurens van der; POSTMA, Eric; HERIK, Jaap van den. Dimensionality reduction: a comparative. **The Journal of Machine Learning Research**, v. 10, n. 66-71, p. 13, 2009. Cit. on pp. 18, 33–35.

MADLA, Christine M.; GAVINS, Francesca K.H.; MERCHANT, Hamid A.; ORLU, Mine; MURDAN, Sudaxshina; BASIT, Abdul W. Let's talk about sex: Differences in drug therapy in males and females. **Advanced Drug Delivery Reviews**, v. 175, p. 113804, Aug. 2021. ISSN 0169409X. DOI: 10.1016/j.addr.2021.05.014. Available from: <https://linkinghub.elsevier.com/retrieve/pii/S0169409X21001782>. Visited on: 22 Jan. 2022. Cit. on p. 18.

MADUGULA, Sita Sirisha; JOHN, Lijo; NAGAMANI, Selvaraman; GAUR, Anamika Singh; POROIKOV, Vladimir V.; SASTRY, G. Narahari. Molecular descriptor analysis of approved drugs using unsupervised learning for drug repurposing. **Computers in Biology and Medicine**, v. 138, p. 104856, Nov. 2021. ISSN 00104825. DOI: 10.1016/j.compbiomed.2021.104856. Available from: <https://linkinghub.elsevier.com/retrieve/pii/S0010482521006508>. Visited on: 19 Feb. 2022. Cit. on pp. 25, 26.

MAHMUD, S.M. Hasan; CHEN, Wenyu; JAHAN, Hosney; LIU, Yougsheng; HASAN, S.M. Mamun. Dimensionality reduction based multi-kernel framework for drug-target interaction prediction. **Chemometrics and Intelligent Laboratory Systems**, v. 212, p. 104270, May 2021. ISSN 01697439. DOI: 10.1016/j.chemolab.2021.104270. Available from: <https://linkinghub.elsevier.com/retrieve/pii/S0169743921000381>. Visited on: 8 Feb. 2022. Cit. on p. 24.

MCCOUBREY, Laura E.; GAISFORD, Simon; ORLU, Mine; BASIT, Abdul W. Predicting drug-microbiome interactions with machine learning. **Biotechnology Advances**, v. 54, p. 107797, Jan. 2022. ISSN 07349750. DOI: 10.1016/j.biotechadv.2021.107797. Available from: <https://linkinghub.elsevier.com/retrieve/pii/S0734975021001038>. Visited on: 23 Jan. 2022. Cit. on p. 23.

MCINNES, Leland; HEALY, John; MELVILLE, James. Umap: Uniform manifold approximation and projection for dimension reduction. **arXiv preprint arXiv:1802.03426**, 2018. Cit. on pp. 35, 37, 38, 50.

MCKINNEY, Wes et al. Data structures for statistical computing in python. In: AUSTIN, TX, 1. PROCEEDINGS of the 9th Python in Science Conference. [S.l.: s.n.], 2010. v. 445, p. 51–56. Cit. on p. 46.

MEHMOOD, Erum; ANEES, Tayyaba. Challenges and Solutions for Processing Real-Time Big Data Stream: A Systematic Literature Review. en. **IEEE Access**, v. 8, p. 119123–119143, 2020. ISSN 2169-3536. DOI: 10.1109/ACCESS.2020.3005268. Available

from: <https://ieeexplore.ieee.org/document/9126812/>. Visited on: 4 Apr. 2021. Cit. on p. 22.

MILOŠEVIĆ, Djuradj; MEDEIROS, Andrew S.; STOJKOVIĆ PIPERAC, Milica; CVIJANOVIĆ, Dušanka; SOININEN, Janne; MILOSAVLJEVIĆ, Aleksandar; PREDIĆ, Bratislav. The application of Uniform Manifold Approximation and Projection (UMAP) for unconstrained ordination and classification of biological indicators in aquatic ecology. **Science of The Total Environment**, v. 815, p. 152365, Apr. 2022. ISSN 00489697. DOI: 10.1016/j.scitotenv.2021.152365. Available from: <https://linkinghub.elsevier.com/retrieve/pii/S0048969721074428>. Visited on: 25 Jan. 2022. Cit. on p. 38.

MORITA, Y.; REZAEIRAVESH, S.; TABATABAEI, N.; VINUESA, R.; FUKAGATA, K.; SCHLATTER, P. Applying Bayesian optimization with Gaussian process regression to computational fluid dynamics problems. **Journal of Computational Physics**, v. 449, p. 110788, Jan. 2022. ISSN 00219991. DOI: 10.1016/j.jcp.2021.110788. Available from: <https://linkinghub.elsevier.com/retrieve/pii/S0021999121006835>. Visited on: 13 Feb. 2022. Cit. on p. 42.

MOZAFARI, Zeinab; CHAMJANGALI, Mansour Arab; ARASHI, Mohammad; GOUDARZI, Nasser. Application of the LAD-LASSO as a dimensional reduction technique in the ANN-based QSAR study: Discovery of potent inhibitors using molecular docking simulation. **Chemometrics and Intelligent Laboratory Systems**, v. 222, p. 104510, Mar. 2022. ISSN 01697439. DOI: 10.1016/j.chemolab.2022.104510. Available from: <https://linkinghub.elsevier.com/retrieve/pii/S0169743922000211>. Visited on: 8 Feb. 2022. Cit. on p. 24.

NAHID, Abdullah-Al; KONG, Yinan. Involvement of Machine Learning for Breast Cancer Image Classification: A Survey. **Computational and Mathematical Methods in Medicine**, 2017. Cit. on p. 34.

NAYARISSERI, Anuraj; KHANDELWAL, Ravina; TANWAR, Poonam; MADHAVI, Maddala; SHARMA, Diksha; THAKUR, Garima; SPECK-PLANCHE, Alejandro; SINGH, Sanjeev Kumar. Artificial Intelligence, Big Data and Machine Learning Approaches in Precision Medicine & Drug Discovery. **Current Drug Targets**, v. 22, n. 6, p. 631–655, 23 Apr. 2021. ISSN 13894501. DOI: 10.2174/1389450122999210104205732. Available from: <https://www.eurekaselect.com/189908/article>. Visited on: 23 Jan. 2022. Cit. on p. 18.

NAZIR, Shah; KHAN, Sulaiman; KHAN, Habib Ullah; ALI, Shaukat; GARCIA-MAGARINO, Ivan; ATAN, Rodziah Binti; NAWAZ, Muhammad. A Comprehensive Analysis of Healthcare Big Data Management, Analytics and Scientific Programming. en. **IEEE Access**, v. 8, p. 95714–95733, 2020. ISSN 2169-3536. DOI:

10.1109/ACCESS.2020.2995572. Available from: <https://ieeexplore.ieee.org/document/9096305/>. Visited on: 4 Apr. 2021. Cit. on pp. 17, 23.

NWEKE, Henry Friday; TEH, Ying Wah; AL-GARADI, Mohammed Ali; ALO, Uzoma Rita. Deep learning algorithms for human activity recognition using mobile and wearable sensor networks: State of the art and research challenges. **Expert Systems with Applications**, v. 105, p. 233–261, 2018. Cit. on p. 34.

OLSON, C.C.; JUDD, K.P.; NICHOLS, J.M. Manifold learning techniques for unsupervised anomaly detection. **Expert Systems with Applications**, v. 91, p. 374–385, 2018. Cit. on pp. 34, 35.

OMAR, Tamer; KETSEOGLOU, Thomas; NAFFAA, Ibrahim. A novel self-healing model using precoding & big-data based approach for 5G networks. en. **Pervasive and Mobile Computing**, v. 73, p. 101365, June 2021. ISSN 15741192. DOI: 10.1016/j.pmcj.2021.101365. Available from: <https://linkinghub.elsevier.com/retrieve/pii/S1574119221000353>. Visited on: 4 Apr. 2021. Cit. on p. 23.

ORSENIGO, Carlotta; VERCELLIS, Carlo. A comparative study of nonlinear manifold learning methods for cancer microarray data classification. **Expert Systems with Applications**, v. 40, n. 6, p. 2189–2197, 2013. Cit. on p. 35.

PATEL, Veer; SHAH, Manan. A comprehensive study on artificial intelligence and machine learning in drug discovery and drug development. **Intelligent Medicine**, s2667102621001066, Nov. 2021. ISSN 26671026. DOI: 10.1016/j.imed.2021.10.001. Available from: <https://linkinghub.elsevier.com/retrieve/pii/S2667102621001066>. Visited on: 22 Jan. 2022. Cit. on p. 22.

PATHAK, Shreyans; PATHAK, Shashwat. Data Visualization Techniques, Model and Taxonomy. In: [s.l.: s.n.], Jan. 2020. P. 249–271. Cit. on p. 23.

PATHIRAGE, Chathurdara Sri Nadith; LI, Jun; LI, Ling; HAO, Hong; LIU, Wanquan; NI, Pinghe. Structural damage identification based on autoencoder neural networks and deep learning. **Engineering Structures**, v. 172, p. 13–28, Oct. 2018. ISSN 01410296. DOI: 10.1016/j.engstruct.2018.05.109. Available from: <https://linkinghub.elsevier.com/retrieve/pii/S0141029618302062>. Visited on: 25 Jan. 2022. Cit. on p. 36.

PEARSON, Karl. On lines and planes of closest fit to systems of points in space. **The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science**, v. 2, n. 11, p. 559–572, Nov. 1901. ISSN 1941-5982, 1941-5990. DOI: 10.1080/14786440109462720. Available from:

<https://www.tandfonline.com/doi/full/10.1080/14786440109462720>. Visited on: 12 Feb. 2022. Cit. on pp. 35, 36.

PEDREGOSA, F.; VAROQUAUX, G.; GRAMFORT, A.; MICHEL, V.; THIRION, B.; GRISEL, O.; BLONDEL, M.; PRETTENHOFER, P.; WEISS, R.; DUBOURG, V.; VANDERPLAS, J.; PASSOS, A.; COURNAPEAU, D.; BRUCHER, M.; PERROT, M.; DUCHESNAY, E. Scikit-learn: Machine Learning in Python. **Journal of Machine Learning Research**, v. 12, p. 2825–2830, 2011. Cit. on p. 48.

PEREIRA, Dulce G.; AFONSO, Anabela; MEDEIROS, Fátima Melo. Overview of Friedman's Test and Post-hoc Analysis. **Communications in Statistics - Simulation and Computation**, v. 44, n. 10, p. 2636–2653, 26 Nov. 2015. ISSN 0361-0918, 1532-4141. DOI: 10.1080/03610918.2014.931971. Available from: <http://www.tandfonline.com/doi/full/10.1080/03610918.2014.931971>. Visited on: 19 Feb. 2022. Cit. on p. 44.

PEREIRA, Rafael B.; PLASTINO, Alexandre; ZADROZNY, Bianca; MERSCHMANN, Luiz H. C. Categorizing feature selection methods for multi-label classification. **Artificial Intelligence Review**, v. 49, n. 1, p. 57–78, Jan. 2018. Cit. on p. 34.

PILARIO, Karl Ezra; TIELEMANS, Alexander; MOJICA, Elmer-Rico E. Geographical discrimination of propolis using dynamic time warping kernel principal components analysis. **Expert Systems with Applications**, v. 187, p. 115938, Jan. 2022. ISSN 09574174. DOI: 10.1016/j.eswa.2021.115938. Available from: <https://linkinghub.elsevier.com/retrieve/pii/S0957417421012926>. Visited on: 25 Jan. 2022. Cit. on p. 37.

POPOVIC, D. Intelligent Control with Neural Networks. In: SOFT Computing and Intelligent Systems. [S.l.]: Elsevier, 2000. P. 419–467. ISBN 978-0-12-646490-0. DOI: 10.1016/B978-012646490-0/50021-4. Available from: <https://linkinghub.elsevier.com/retrieve/pii/B9780126464900500214>. Visited on: 13 Feb. 2022. Cit. on p. 40.

RACE, Alan M.; SUTTON, Daniel; HAMM, Gregory; MAGLENNON, Gareth; MORTON, Jennifer P.; STRITTMATTER, Nicole; CAMPBELL, Andrew; SANSOM, Owen J.; WANG, Yinhai; BARRY, Simon T.; TAKÁTS, Zoltan; GOODWIN, Richard J. A.; BUNCH, Josephine. Deep Learning-Based Annotation Transfer between Molecular Imaging Modalities: An Automated Workflow for Multimodal Data Integration. **Analytical Chemistry**, v. 93, n. 6, p. 3061–3071, 16 Feb. 2021. ISSN 0003-2700, 1520-6882. DOI: 10.1021/acs.analchem.0c02726. Available from: <https://pubs.acs.org/doi/10.1021/acs.analchem.0c02726>. Visited on: 25 Jan. 2022. Cit. on p. 38.

REHMAN, Muhammad Habib ur; YAQOOB, Ibrar; SALAH, Khaled; IMRAN, Muhammad; JAYARAMAN, Prem Prakash; PERERA, Charith. The role of big data analytics in industrial Internet of Things. **Future Generation Computer Systems**, v. 99, p. 247–259, 2019. Cit. on p. 22.

ROBINSON, R.; HAVILAND, J.S. Understanding Statistical Significance and Avoiding Common Pitfalls. **Clinical Oncology**, v. 33, n. 12, p. 804–806, Dec. 2021. ISSN 09366555. DOI: 10.1016/j.clon.2021.06.008. Available from: <https://linkinghub.elsevier.com/retrieve/pii/S0936655521002284>. Visited on: 14 Feb. 2022. Cit. on p. 45.

RODRIGUES, Daniele B; FAILLA, Mark L. Intestinal cell models for investigating the uptake, metabolism and absorption of dietary nutrients and bioactive compounds. **Current Opinion in Food Science**, v. 41, p. 169–179, Oct. 2021. ISSN 22147993. DOI: 10.1016/j.cofs.2021.04.002. Available from: <https://linkinghub.elsevier.com/retrieve/pii/S2214799321000680>. Visited on: 24 Jan. 2022. Cit. on p. 30.

ROSENBAUM, Sara. **Basic pharmacokinetics and pharmacodynamics: an integrated textbook and computer simulations**. [S.l.]: Wiley, 2017. v. 2. Cit. on p. 18.

ROWEIS, Sam T.; SAUL, Lawrence K. Nonlinear Dimensionality Reduction by Locally Linear Embedding. **Science**, American Association for the Advancement of Science, v. 290, n. 5500, p. 2323–2326, 2000. Cit. on p. 34.

RUSSELL, Stuart; NORVIG, Peter. **Artificial Intelligence: A Modern Approach**. 4. ed. [S.l.]: Pearson, 2021. (Pearson series in artificial intelligence). ISBN 9781292401133. Available from: <https://books.google.com.br/books?id=koFptAEACAAJ>. Cit. on pp. 32, 33, 39–41, 43.

SANTANA, Kauê; NASCIMENTO, Lidiane Diniz do; LIMA E LIMA, Anderson; DAMASCENO, Vinícius; NAHUM, Claudio; BRAGA, Rodolpho C.; LAMEIRA, Jerônimo. Applications of Virtual Screening in Bioprospecting: Facts, Shifts, and Perspectives to Explore the Chemo-Structural Diversity of Natural Products. **Frontiers in Chemistry**, v. 9, p. 662688, 29 Apr. 2021. ISSN 2296-2646. DOI: 10.3389/fchem.2021.662688. Available from: <https://www.frontiersin.org/articles/10.3389/fchem.2021.662688/full>. Visited on: 28 Feb. 2022. Cit. on p. 31.

SANTOS, Reginaldo; BORGES, Gilvan; SANTOS, Adam; SILVA, Moisés; SALES, Claudomiro; COSTA, João C.W. A. A rotationally invariant semi-autonomous particle swarm optimizer with directional diversity. **Swarm and Evolutionary Computation**, v. 56, p. 100700, Aug. 2020. ISSN 22106502. DOI: 10.1016/j.swevo.2020.100700. Available from:

<https://linkinghub.elsevier.com/retrieve/pii/S2210650219305504>. Visited on: 21 May 2021. Cit. on p. 44.

SANTOS, Reginaldo; BORGES, Gilvan; SANTOS, Adam; SILVA, Moisés; SALES, Claudomiro; COSTA, João C.W.A. Empirical study on rotation and information exchange in particle swarm optimization. **Swarm and Evolutionary Computation**, v. 48, p. 312–328, 2019. Cit. on p. 44.

SCHAFFER, Cullen. Selecting a classification method by cross-validation. **Machine Learning**, v. 13, n. 1, p. 135–143, Oct. 1993. ISSN 0885-6125, 1573-0565. DOI: 10.1007/BF00993106. Available from: <http://link.springer.com/10.1007/BF00993106>. Visited on: 13 Feb. 2022. Cit. on p. 34.

SCHÖLKOPF, Bernhard; SMOLA, Alexander; MÜLLER, Klaus-Robert. Kernel principal component analysis. In: GERSTNER, Wulfram; GERMOND, Alain; HASLER, Martin; NICOUD, Jean-Daniel (Eds.). **Artificial Neural Networks — ICANN'97**. Red. by Gerhard Goos, Juris Hartmanis and Jan van Leeuwen. Berlin, Heidelberg: Springer Berlin Heidelberg, 1997. v. 1327. Series Title: Lecture Notes in Computer Science. P. 583–588. ISBN 978-3-540-69620-9. DOI: 10.1007/BFb0020217. Available from: <http://link.springer.com/10.1007/BFb0020217>. Visited on: 25 Jan. 2022. Cit. on pp. 35, 37.

SEWELL, Daniel K. Visualizing data through curvilinear representations of matrices. **Computational Statistics & Data Analysis**, v. 128, p. 255–270, 2018. ISSN 0167-9473. Cit. on pp. 18, 22, 23.

SHARIFZADEH, Sara; GHODSI, Ali; CLEMMENSEN, Line H.; ERSBØLL, Bjarne K. Sparse supervised principal component analysis (SSPCA) for dimension reduction and variable selection. **Engineering Applications of Artificial Intelligence**, v. 65, p. 168–177, 2017. ISSN 0952-1976. Cit. on p. 35.

SHIN, Moonshik; JANG, Donjin; NAM, Hojung; LEE, Kwang Hyung; LEE, Doheon. Predicting the Absorption Potential of Chemical Compounds Through a Deep Learning Approach. **IEEE/ACM Transactions on Computational Biology and Bioinformatics**, v. 15, n. 2, p. 432–440, Mar. 2018. Conference Name: IEEE/ACM Transactions on Computational Biology and Bioinformatics. ISSN 1557-9964. DOI: 10.1109/TCBB.2016.2535233. Cit. on p. 25.

SINGH, Kunwar P.; GUPTA, Shikha; BASANT, Nikita. In silico prediction of cellular permeability of diverse chemicals using qualitative and quantitative SAR modeling approaches. **Chemometrics and Intelligent Laboratory Systems**, v. 140, p. 61–72, Jan. 2015. ISSN 01697439. DOI: 10.1016/j.chemolab.2014.10.005. Available from: <https://linkinghub.elsevier.com/retrieve/pii/S0169743914002135>. Visited on: 9 Feb. 2022. Cit. on pp. 20, 24.

SIVARAJAH, Uthayasankar; KAMAL, Muhammad Mustafa; IRANI, Zahir; WEERAKKODY, Vishanth. Critical analysis of Big Data challenges and analytical methods. **Journal of Business Research**, v. 70, p. 263–286, 2017. Cit. on p. 22.

SNOEK, Jasper; LAROCHELLE, Hugo; ADAMS, Ryan P. Practical Bayesian Optimization of Machine Learning Algorithms. In: ADVANCES in Neural Information Processing Systems. [S.l.]: Curran Associates, Inc., 2012. v. 25. Available from: <https://proceedings.neurips.cc/paper/2012/hash/05311655a15b75fab86956663e1819cd-Abstract.html>. Visited on: 13 Feb. 2022. Cit. on pp. 41, 42.

SONG, Juyoung; HAN, Yoonsun; KIM, Kwanghyun; SONG, Tae Min. Social big data analysis of future signals for bullying in South Korea: Application of general strain theory. en. **Telematics and Informatics**, v. 54, p. 101472, Nov. 2020. ISSN 07365853. DOI: 10.1016/j.tele.2020.101472. Available from: <https://linkinghub.elsevier.com/retrieve/pii/S0736585320301313>. Visited on: 4 Apr. 2021. Cit. on p. 22.

SONG, M.; YANG, H.; SIADAT, S.H.; PECHENIZKIY, M. A comparative study of dimensionality reduction techniques to enhance trace clustering performances. **Expert Systems with Applications**, v. 40, n. 9, p. 3722–3737, 2013. Cit. on pp. 18, 34.

STORPIRTIS, S.; AL., et. **Farmacocinética Básica e Aplicada**. [S.l.]: GEN, 2011. v. 1. Cit. on pp. 18, 29.

SZUBERT, Benjamin; COLE, Jennifer E; MONACO, Claudia; DROZDOV, Ignat. Structure-preserving visualisation of high dimensional single-cell datasets. **Scientific reports**, Nature Publishing Group, v. 9, n. 1, p. 1–10, 2019. Cit. on pp. 17, 35, 36, 38, 39, 50.

TA, Giang Huong; JHANG, Cin-Syong; WENG, Ching-Feng; LEONG, Max K. Development of a Hierarchical Support Vector Regression-Based In Silico Model for Caco-2 Permeability. **Pharmaceutics**, v. 13, n. 2, p. 174, 28 Jan. 2021. ISSN 1999-4923. DOI: 10.3390/pharmaceutics13020174. Available from: <https://www.mdpi.com/1999-4923/13/2/174>. Visited on: 8 Feb. 2022. Cit. on p. 25.

TERPILOWSKI, Maksim A. scikit-posthocs: Pairwise multiple comparison tests in Python. **Journal of Open Source Software**, v. 4, n. 36, p. 1169, 2019. Cit. on p. 52.

TIAN, Hao; TAO, Peng. ivis Dimensionality Reduction Framework for Biomacromolecular Simulations. **Journal of Chemical Information and Modeling**, v. 60, n. 10, p. 4569–4581, 26 Oct. 2020. ISSN 1549-9596, 1549-960X. DOI: 10.1021/acs.jcim.0c00485. Available from: <https://pubs.acs.org/doi/10.1021/acs.jcim.0c00485>. Visited on: 24 Jan. 2022. Cit. on pp. 35, 38, 39.

TOZER, T. N.; ROWLAN, M. **Essentials of Pharmacokinetics and Pharmacodynamics**. [S.l.]: Wolters Kluwer, 2016. v. 2. Cit. on pp. 21, 28.

TSATSISHVILI, Valeri; BURUNAT, Iballa; CONG, Fengyu; TOIVIAINEN, Petri; ALLURI, Vinoo; RISTANIEMI, Tapani. On application of kernel PCA for generating stimulus features for fMRI during continuous music listening. **Journal of Neuroscience Methods**, v. 303, p. 1–6, June 2018. ISSN 01650270. DOI: 10.1016/j.jneumeth.2018.03.014. Available from: <https://linkinghub.elsevier.com/retrieve/pii/S0165027018300852>. Visited on: 25 Jan. 2022. Cit. on p. 37.

TURAGA, Pavan; ANIRUDH, Rushil; CHELLAPPA, Rama. Manifold Learning. **Computer Vision: A Reference Guide**, Springer, p. 1–6, 2020. Cit. on p. 35.

VAN DER MAATEN, Laurens; HINTON, Geoffrey. Visualizing data using t-SNE. **Journal of machine learning research**, v. 9, n. 11, 2008. Cit. on p. 38.

VASCONCELOS, Matheus; FLEXA, Caio; MOREIRA, Igor; SANTOS, Reginaldo; SALES, Claudomiro. Improving Particle Swarm Optimization with Self-adaptive Parameters, Rotational Invariance, and Diversity Control. In_____. **Intelligent Systems**. Cham: Springer International Publishing, 2021. P. 218–233. ISBN 978-3-030-91702-9. Cit. on pp. 44, 84.

VEBER, D. F.; AL., et. Molecular properties that influence the oral bioavailability of drug candidates. **J Med Chem**, v. 45, p. 2615–23, 2002. DOI: 10.1021/jm020017n. Cit. on p. 31.

VERMEULEN, Marc; SMITH, Kate; EREMIN, Katherine; RAYNER, Georgina; WALTON, Marc. Application of Uniform Manifold Approximation and Projection (UMAP) in spectral imaging of artworks. **Spectrochimica Acta Part A: Molecular and Biomolecular Spectroscopy**, v. 252, p. 119547, May 2021. ISSN 13861425. DOI: 10.1016/j.saa.2021.119547. Available from: <https://linkinghub.elsevier.com/retrieve/pii/S1386142521001232>. Visited on: 25 Jan. 2022. Cit. on p. 38.

VIJAYAN, R.S.K.; KIHLBERG, Jan; CROSS, Jason B.; POONGAVANAM, Vasanthanathan. Enhancing preclinical drug discovery with artificial intelligence. **Drug Discovery Today**, s1359644621005043, Nov. 2021. ISSN 13596446. DOI: 10.1016/j.drudis.2021.11.023. Available from: <https://linkinghub.elsevier.com/retrieve/pii/S1359644621005043>. Visited on: 22 Jan. 2022. Cit. on pp. 17–20, 29.

VINOTHA, G.; SUNDAR, T.V. Drug Likeness Prediction Using Structure Based Molecular Descriptors and Support Vector Machines. **Materials Today: Proceedings**, v. 18, p. 1658–1669, 2019. ISSN 22147853. DOI: 10.1016/j.matpr.2019.05.262.

Available from:
<https://linkinghub.elsevier.com/retrieve/pii/S2214785319311083>. Visited on: 19 Feb. 2022. Cit. on p. 30.

VIRTANEN, Pauli; GOMMERS, Ralf; OLIPHANT, Travis E.; HABERLAND, Matt; REDDY, Tyler; COURNAPEAU, David; BUROVSKI, Evgeni; PETERSON, Pearu; WECKESSER, Warren; BRIGHT, Jonathan; VAN DER WALT, Stéfan J.; BRETT, Matthew; WILSON, Joshua; MILLMAN, K. Jarrod; MAYOROV, Nikolay; NELSON, Andrew R. J.; JONES, Eric; KERN, Robert; LARSON, Eric; CAREY, C J; POLAT, İlhan; FENG, Yu; MOORE, Eric W.; VANDERPLAS, Jake; LAXALDE, Denis; PERKTOLD, Josef; CIMRMAN, Robert; HENRIKSEN, Ian; QUINTERO, E. A.; HARRIS, Charles R.; ARCHIBALD, Anne M.; RIBEIRO, Antônio H.; PEDREGOSA, Fabian; VAN MULBREGT, Paul; SCIPY 1.0 CONTRIBUTORS. SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python. **Nature Methods**, v. 17, p. 261–272, 2020. DOI: 10.1038/s41592-019-0686-2. Cit. on p. 52.

WAGNER, John G. History of pharmacokinetics. **Pharmacology & Therapeutics**, v. 12, n. 3, p. 537–562, 1981. DOI: 10.1016/0163-7258(81)90097-8. Cit. on p. 29.

WAMBA, Samuel Fosso; AKTER, Shahriar; EDWARDS, Andrew; CHOPIN, Geoffrey; GNANZOU, Denis. How 'big data' can make big impact: Findings from a systematic review and a longitudinal case study. **International Journal of Production Economics**, v. 165, p. 234–246, 2015. Cit. on p. 22.

WAN, Youchuan; WANG, Mingwei; YE, Zhiwei; LAI, Xudong. A Feature Selection Method Based on Modified Binary Coded Ant Colony Optimization Algorithm. **Appl. Soft Comput.**, Elsevier Science Publishers B. V., Amsterdam, The Netherlands, The Netherlands, v. 49, n. 100, p. 248–258, Dec. 2016. ISSN 1568-4946. Cit. on p. 34.

WANG, Fei; SUN, Jimeng. Survey on distance metric learning and dimensionality reduction in data mining. **Data Mining and Knowledge Discovery**, v. 29, n. 2, p. 534–564, Mar. 2015. Cit. on pp. 18, 33–35.

WANG, Ke; DOWLING, Alexander W. Bayesian optimization for chemical products and functional materials. **Current Opinion in Chemical Engineering**, v. 36, p. 100728, June 2022. ISSN 22113398. DOI: 10.1016/j.coche.2021.100728. Available from: <https://linkinghub.elsevier.com/retrieve/pii/S2211339821000605>. Visited on: 13 Feb. 2022. Cit. on p. 42.

WANG, Ning-Ning; HUANG, Chen; DONG, Jie; YAO, Zhi-Jiang; ZHU, Min-Feng; DENG, Zhen-Ke; LV, Ben; LU, Ai-Ping; CHEN, Alex F.; CAO, Dong-Sheng. Predicting human intestinal absorption with modified random forest approach: a comprehensive evaluation of molecular representation, unbalanced data, and applicability domain issues. **RSC Advances**, v. 7, n. 31, p. 19007–19018, 2017. ISSN 2046-2069. DOI:

10.1039/C6RA28442F. Available from: <http://xlink.rsc.org/?DOI=C6RA28442F>. Visited on: 2 Feb. 2022. Cit. on pp. 21, 24, 40, 48.

WANG, Wei; YE, Zhuyifan; GAO, Hanlu; OUYANG, Defang. Computational pharmaceutics - A new paradigm of drug delivery. **Journal of Controlled Release**, v. 338, p. 119–136, Oct. 2021. ISSN 01683659. DOI: 10.1016/j.jconrel.2021.08.030. Available from: <https://linkinghub.elsevier.com/retrieve/pii/S0168365921004363>. Visited on: 22 Jan. 2022. Cit. on p. 17.

WARING, Jonathan; LINDVALL, Charlotta; UMETON, Renato. Automated machine learning: Review of the state-of-the-art and opportunities for healthcare. **Artificial Intelligence in Medicine**, v. 104, p. 101822, Apr. 2020. ISSN 09333657. DOI: 10.1016/j.artmed.2020.101822. Available from: <https://linkinghub.elsevier.com/retrieve/pii/S0933365719310437>. Visited on: 9 Feb. 2022. Cit. on p. 42.

WASKOM, Michael L. seaborn: statistical data visualization. **Journal of Open Source Software**, The Open Journal, v. 6, n. 60, p. 3021, 2021. DOI: 10.21105/joss.03021. Available from: <https://doi.org/10.21105/joss.03021>. Cit. on p. 46.

WERBOS, Paul. **Beyond regression: new tools for prediction and analysis in the behavioral sciences**. 1974. PhD thesis – Harvard University. Cit. on p. 40.

WOLPERT, D.H.; MACREADY, W.G. No free lunch theorems for optimization. **IEEE Transactions on Evolutionary Computation**, v. 1, n. 1, p. 67–82, Apr. 1997. ISSN 1089778X. DOI: 10.1109/4235.585893. Available from: <http://ieeexplore.ieee.org/document/585893/>. Visited on: 9 Feb. 2022. Cit. on p. 33.

WOLPERT, David H. The Lack of A Priori Distinctions Between Learning Algorithms. **Neural Computation**, v. 8, n. 7, p. 1341–1390, Oct. 1996. ISSN 0899-7667, 1530-888X. DOI: 10.1162/neco.1996.8.7.1341. Available from: <https://direct.mit.edu/neco/article/8/7/1341-1390/6016>. Visited on: 9 Feb. 2022. Cit. on p. 33.

XU, Peiliang. Truncated SVD methods for discrete linear ill-posed problems. **Geophysical Journal International**, v. 135, n. 2, p. 505–514, Nov. 1998. ISSN 0956540X, 1365246X. DOI: 10.1046/j.1365-246X.1998.00652.x. Available from: <https://academic.oup.com/gji/article-lookup/doi/10.1046/j.1365-246X.1998.00652.x>. Visited on: 10 Feb. 2022. Cit. on p. 36.

XU, Yan; SUN, Yanming; WAN, Jiafu; LIU, Xiaolong; SONG, Zhiting. Industrial Big Data for Fault Diagnosis: Taxonomy, Review, and Applications. en. **IEEE Access**, v. 5, p. 17368–17380, 2017. ISSN 2169-3536. DOI: 10.1109/ACCESS.2017.2731945. Available

from: <http://ieeexplore.ieee.org/document/7990488/>. Visited on: 4 Apr. 2021. Cit. on p. 22.

YANG, Li; SHAMI, Abdallah. On hyperparameter optimization of machine learning algorithms: Theory and practice. **Neurocomputing**, v. 415, p. 295–316, Nov. 2020. ISSN 09252312. DOI: 10.1016/j.neucom.2020.07.061. Available from: <https://linkinghub.elsevier.com/retrieve/pii/S0925231220311693>. Visited on: 13 Feb. 2022. Cit. on pp. 23, 34, 41, 42, 49.

YANG, Ming; CHEN, Jialei; XU, Liwen; SHI, Xiufeng; ZHOU, Xin; XI, Zhijun; AN, Rui; WANG, Xinhong. A novel adaptive ensemble classification framework for ADME prediction. **RSC Advances**, v. 8, n. 21, p. 11661–11683, 2018. ISSN 2046-2069. DOI: 10.1039/C8RA01206G. Available from: <http://xlink.rsc.org/?DOI=C8RA01206G>. Visited on: 29 Jan. 2022. Cit. on p. 18.

YANG, Yang; SUN, Hongjian; ZHANG, Yu; ZHANG, Tiefu; GONG, Jialei; WEI, Yunbo; DUAN, Yong-Gang; SHU, Minglei; YANG, Yuchen; WU, Di; YU, Di. Dimensionality reduction by UMAP reinforces sample heterogeneity analysis in bulk transcriptomic data. **Cell Reports**, v. 36, n. 4, p. 109442, July 2021. ISSN 22111247. DOI: 10.1016/j.celrep.2021.109442. Available from: <https://linkinghub.elsevier.com/retrieve/pii/S2211124721008597>. Visited on: 14 Feb. 2022. Cit. on p. 38.

YAQOOB, Ibrar; HASHEM, Ibrahim Abaker Targio; GANI, Abdullah; MOKHTAR, Salimah; AHMED, Ejaz; ANUAR, Nor Badrul; VASILAKOS, Athanasios V. Big data: From beginning to future. **International Journal of Information Management**, v. 36, 6, Part B, p. 1231–1247, 2016. ISSN 0268-4012. Cit. on p. 22.

YOUHANNA, Sonia; LAUSCHKE, Volker M. The Past, Present and Future of Intestinal In Vitro Cell Systems for Drug Absorption Studies. **Journal of Pharmaceutical Sciences**, v. 110, n. 1, p. 50–65, Jan. 2021. ISSN 00223549. DOI: 10.1016/j.xphs.2020.07.001. Available from: <https://linkinghub.elsevier.com/retrieve/pii/S0022354920303622>. Visited on: 22 Jan. 2022. Cit. on pp. 17, 18, 21, 28, 30.

ZHANG, Songtao; DUBAY, Rickey; CHAREST, Meaghan. A principal component analysis model-based predictive controller for controlling part warpage in plastic injection molding. **Expert Systems with Applications**, v. 42, n. 6, p. 2919–2927, 2015. Cit. on p. 36.

ZHAO, Yuan H.; LE, Joelle; ABRAHAM, Michael H.; HERSEY, Anne; EDDERSHAW, Peter J.; LUSCOMBE, Chris N.; BOUTINA, Darko; BECK, Gordon; SHERBORNE, Brad; COOPER, Ian; PLATTS, James A. Evaluation of human intestinal absorption data and subsequent derivation of a quantitative structure–activity relationship (QSAR) with the Abraham descriptors. **Journal of Pharmaceutical Sciences**, v. 90,

n. 6, p. 749–784, June 2001. ISSN 00223549. DOI: 10.1002/jps.1031. Available from: <https://linkinghub.elsevier.com/retrieve/pii/S0022354916307651>. Visited on: 8 Feb. 2022. Cit. on pp. 20, 31.

ZHONG, Xiao; ENKE, David. Forecasting daily stock market return using dimensionality reduction. **Expert Systems with Applications**, v. 67, p. 126–139, 2017. Cit. on p. 36.

ZHOU, Liang; WEISKOPF, Daniel. Multivariate visualization of particle data. **The European Physical Journal Special Topics**, v. 227, n. 14, p. 1741–1755, Mar. 2019. Cit. on p. 35.

ŽILINSKAS, Antanas. Visualization of a statistical approximation of the Pareto front. **Applied Mathematics and Computation**, v. 271, p. 694–700, 2015. ISSN 0096-3003. Cit. on p. 17.

ZOU, Quan; QU, Kaiyang; LUO, Yamei; YIN, Dehui; JU, Ying; TANG, Hua. Predicting Diabetes Mellitus With Machine Learning Techniques. **Frontiers in Genetics**, v. 9, p. 515, 6 Nov. 2018. ISSN 1664-8021. DOI: 10.3389/fgene.2018.00515. Available from: <https://www.frontiersin.org/article/10.3389/fgene.2018.00515/full>. Visited on: 25 Jan. 2022. Cit. on p. 36.

APPENDIX

A

CROSS-VALIDATION MEASUREMENTS PER FOLD

Supplementing the mean CV accuracy scores table exposed in Chapter 4, this table exposes the train and test scores obtained in each fold. Due to implementation limitations, the same could not be made for the fit and predict times.

Table 11 – Accuracy scores of best-performing pipelines in 10-fold stratified CV. Source: the author.

			Measure	Train accuracy	Test accuracy
Dimensionality	Projector	Classifier	Fold		
50	—	KNN	0	0.747368	0.500000
			1	0.715789	0.500000
			2	0.748691	0.761905
			3	0.743455	0.666667
			4	0.717277	0.809524
			5	0.738220	0.761905
			6	0.759162	0.571429
			7	0.748691	0.761905
			8	0.764398	0.619048
			9	0.722513	0.619048
		MLP	0	0.500000	0.500000
			1	0.500000	0.500000
			2	0.502618	0.476190
			3	0.502618	0.476190

Continued on next page

Table 11 – Accuracy of best-performing pipelines in 10-fold stratified CV. Source: the author.

			Measure	Train accuracy	Test accuracy
Dimensionality	Projector	Classifier	Fold		
			4	0.502618	0.476190
			5	0.502618	0.476190
			6	0.502618	0.476190
			7	0.502618	0.476190
			8	0.502618	0.476190
			9	0.502618	0.476190
		SVM	0	0.926316	0.863636
			1	0.910526	0.681818
			2	0.916230	0.809524
			3	0.931937	0.761905
			4	0.895288	0.857143
			5	0.921466	0.714286
			6	0.910995	0.714286
			7	0.937173	0.809524
			8	0.905759	0.904762
			9	0.921466	0.666667
		RF	0	1.000000	0.909091
			1	1.000000	0.772727
			2	1.000000	0.857143
			3	1.000000	0.761905
			4	0.994764	0.904762
			5	0.994764	0.857143
			6	1.000000	0.714286
			7	1.000000	0.809524
			8	1.000000	0.904762
			9	1.000000	0.761905
2	PCA	KNN	0	1.000000	0.772727
			1	1.000000	0.590909
			2	1.000000	0.619048
			3	1.000000	0.761905
			4	1.000000	0.809524

Continued on next page

Table 11 – Accuracy of best-performing pipelines in 10-fold stratified CV. Source: the author.

			Measure	Train accuracy	Test accuracy
Dimensionality	Projector	Classifier	Fold		
			5	1.000000	0.761905
			6	1.000000	0.666667
			7	1.000000	0.666667
			8	1.000000	0.714286
			9	1.000000	0.714286
		MLP	0	0.736842	0.772727
			1	0.747368	0.681818
			2	0.685864	0.666667
			3	0.696335	0.761905
			4	0.712042	0.857143
			5	0.675393	0.619048
			6	0.696335	0.619048
			7	0.701571	0.666667
			8	0.675393	0.666667
			9	0.732984	0.619048
		SVM	0	0.810526	0.772727
			1	0.805263	0.545455
			2	0.774869	0.714286
			3	0.811518	0.666667
			4	0.764398	0.761905
			5	0.790576	0.714286
			6	0.806283	0.666667
			7	0.801047	0.619048
			8	0.795812	0.714286
			9	0.821990	0.666667
		RF	0	0.784211	0.818182
			1	0.810526	0.636364
			2	0.743455	0.809524
			3	0.795812	0.761905
			4	0.774869	0.857143
			5	0.785340	0.761905

Continued on next page

Table 11 – Accuracy of best-performing pipelines in 10-fold stratified CV. Source: the author.

			Measure	Train accuracy	Test accuracy
Dimensionality	Projector	Classifier	Fold		
			6	0.795812	0.714286
			7	0.801047	0.666667
			8	0.785340	0.714286
			9	0.790576	0.714286
	TSVD	KNN	0	1.000000	0.727273
			1	1.000000	0.727273
			2	1.000000	0.666667
			3	1.000000	0.571429
			4	1.000000	0.714286
			5	1.000000	0.571429
			6	1.000000	0.714286
			7	1.000000	0.666667
			8	1.000000	0.761905
			9	1.000000	0.476190
		MLP	0	0.636842	0.681818
			1	0.684211	0.500000
			2	0.685864	0.666667
			3	0.654450	0.761905
			4	0.670157	0.714286
			5	0.628272	0.666667
			6	0.670157	0.714286
			7	0.675393	0.619048
			8	0.670157	0.666667
			9	0.664921	0.571429
		SVM	0	0.821053	0.681818
			1	0.826316	0.727273
			2	0.795812	0.619048
			3	0.853403	0.523810
			4	0.801047	0.761905
			5	0.827225	0.666667
			6	0.806283	0.666667

Continued on next page

Table 11 – Accuracy of best-performing pipelines in 10-fold stratified CV. Source: the author.

			Measure	Train accuracy	Test accuracy
Dimensionality	Projector	Classifier	Fold		
			7	0.874346	0.571429
			8	0.780105	0.761905
			9	0.863874	0.476190
		RF	0	1.000000	0.636364
			1	1.000000	0.681818
			2	1.000000	0.666667
			3	1.000000	0.619048
			4	1.000000	0.761905
			5	1.000000	0.619048
			6	1.000000	0.666667
			7	1.000000	0.666667
			8	1.000000	0.761905
			9	1.000000	0.714286
	PCS	KNN	0	0.710526	0.681818
			1	0.684211	0.681818
			2	0.727749	0.476190
			3	0.643979	0.952381
			4	0.717277	0.476190
			5	0.712042	0.476190
			6	0.691099	0.714286
			7	0.675393	0.666667
			8	0.691099	0.666667
			9	0.701571	0.571429
		MLP	0	0.668421	0.636364
			1	0.668421	0.681818
			2	0.680628	0.619048
			3	0.643979	0.857143
			4	0.654450	0.476190
			5	0.675393	0.476190
			6	0.670157	0.619048
			7	0.670157	0.619048

Continued on next page

Table 11 – Accuracy of best-performing pipelines in 10-fold stratified CV. Source: the author.

			Measure	Train accuracy	Test accuracy
Dimensionality	Projector	Classifier	Fold		
			8	0.664921	0.666667
			9	0.675393	0.571429
		SVM	0	0.715789	0.681818
			1	0.705263	0.727273
			2	0.727749	0.666667
			3	0.675393	0.904762
			4	0.717277	0.476190
			5	0.717277	0.571429
			6	0.696335	0.666667
			7	0.701571	0.666667
			8	0.706806	0.666667
			9	0.743455	0.476190
		RF	0	1.000000	0.636364
			1	1.000000	0.500000
			2	1.000000	0.476190
			3	1.000000	0.619048
			4	1.000000	0.476190
			5	1.000000	0.523810
			6	1.000000	0.666667
			7	1.000000	0.619048
			8	1.000000	0.666667
			9	1.000000	0.571429
	KPCA	KNN	0	0.742105	0.818182
			1	0.763158	0.681818
			2	0.743455	0.761905
			3	0.717277	0.857143
			4	0.764398	0.809524
			5	0.764398	0.714286
			6	0.743455	0.714286
			7	0.743455	0.666667
			8	0.753927	0.714286

Continued on next page

Table 11 – Accuracy of best-performing pipelines in 10-fold stratified CV. Source: the author.

			Measure	Train accuracy	Test accuracy
Dimensionality	Projector	Classifier	Fold		
			9	0.743455	0.619048
		MLP	0	0.689474	0.681818
			1	0.721053	0.500000
			2	0.696335	0.809524
			3	0.701571	0.761905
			4	0.706806	0.809524
			5	0.706806	0.666667
			6	0.680628	0.714286
			7	0.706806	0.666667
			8	0.717277	0.714286
			9	0.727749	0.571429
		SVM	0	0.710526	0.681818
			1	0.731579	0.500000
			2	0.685864	0.809524
			3	0.696335	0.714286
			4	0.696335	0.809524
			5	0.712042	0.761905
			6	0.701571	0.619048
			7	0.712042	0.714286
			8	0.722513	0.761905
			9	0.727749	0.619048
		RF	0	1.000000	0.727273
			1	0.994737	0.590909
			2	1.000000	0.809524
			3	1.000000	0.666667
			4	1.000000	0.761905
			5	1.000000	0.761905
			6	1.000000	0.666667
			7	1.000000	0.761905
			8	1.000000	0.714286
			9	1.000000	0.619048

Continued on next page

Table 11 – Accuracy of best-performing pipelines in 10-fold stratified CV. Source: the author.

			Measure	Train accuracy	Test accuracy
Dimensionality	Projector	Classifier	Fold		
	UMAP	KNN	0	1.000000	0.590909
			1	1.000000	0.636364
			2	1.000000	0.857143
			3	1.000000	0.809524
			4	1.000000	0.809524
			5	1.000000	0.857143
			6	1.000000	0.761905
			7	1.000000	0.761905
			8	1.000000	0.809524
			9	1.000000	0.714286
		MLP	0	0.500000	0.500000
			1	0.500000	0.500000
			2	0.497382	0.523810
			3	0.497382	0.523810
			4	0.497382	0.523810
			5	0.497382	0.523810
			6	0.502618	0.476190
			7	0.502618	0.476190
			8	0.502618	0.476190
			9	0.502618	0.476190
		SVM	0	0.752632	0.590909
			1	0.715789	0.500000
			2	0.549738	0.666667
			3	0.680628	0.809524
			4	0.701571	0.809524
			5	0.816754	0.761905
			6	0.732984	0.571429
			7	0.774869	0.619048
			8	0.738220	0.761905
			9	0.623037	0.476190
		RF	0	1.000000	0.681818

Continued on next page

Table 11 – Accuracy of best-performing pipelines in 10-fold stratified CV. Source: the author.

			Measure	Train accuracy	Test accuracy
Dimensionality	Projector	Classifier	Fold		
			1	1.000000	0.727273
			2	1.000000	0.714286
			3	1.000000	0.809524
			4	1.000000	0.809524
			5	1.000000	0.761905
			6	1.000000	0.809524
			7	1.000000	0.761905
			8	1.000000	0.904762
			9	1.000000	0.761905
	Ivis	KNN	0	1.000000	0.772727
			1	1.000000	0.727273
			2	1.000000	0.904762
			3	1.000000	0.857143
			4	1.000000	0.952381
			5	1.000000	0.857143
			6	1.000000	0.809524
			7	1.000000	0.761905
			8	1.000000	0.857143
			9	1.000000	0.809524
		MLP	0	0.500000	0.500000
			1	0.500000	0.500000
			2	0.502618	0.476190
			3	0.502618	0.476190
			4	0.502618	0.476190
			5	0.502618	0.476190
			6	0.502618	0.476190
			7	0.502618	0.476190
			8	0.502618	0.476190
			9	0.502618	0.476190
		SVM	0	0.984211	0.818182
			1	0.989474	0.727273

Continued on next page

Table 11 – Accuracy of best-performing pipelines in 10-fold stratified CV. Source: the author.

			Measure	Train accuracy	Test accuracy
Dimensionality	Projector	Classifier	Fold		
			2	1.000000	0.904762
			3	1.000000	0.857143
			4	0.989529	1.000000
			5	0.947644	0.809524
			6	1.000000	0.714286
			7	1.000000	0.714286
			8	0.994764	0.857143
			9	1.000000	0.857143
		RF	0	1.000000	0.818182
			1	1.000000	0.772727
			2	1.000000	0.809524
			3	1.000000	0.714286
			4	1.000000	0.952381
			5	1.000000	0.904762
			6	1.000000	0.761905
			7	1.000000	0.761905
			8	1.000000	0.904762
			9	1.000000	0.809524
3	PCA	KNN	0	0.773684	0.818182
			1	0.789474	0.590909
			2	0.774869	0.761905
			3	0.764398	0.809524
			4	0.748691	0.809524
			5	0.764398	0.809524
			6	0.780105	0.666667
			7	0.738220	0.666667
			8	0.764398	0.714286
			9	0.785340	0.666667
		MLP	0	0.731579	0.772727
			1	0.763158	0.636364
			2	0.738220	0.809524

Continued on next page

Table 11 – Accuracy of best-performing pipelines in 10-fold stratified CV. Source: the author.

			Measure	Train accuracy	Test accuracy
Dimensionality	Projector	Classifier	Fold		
			3	0.753927	0.809524
			4	0.732984	0.857143
			5	0.727749	0.809524
			6	0.753927	0.666667
			7	0.738220	0.714286
			8	0.738220	0.666667
			9	0.748691	0.666667
		SVM	0	0.931579	0.727273
			1	0.947368	0.818182
			2	0.931937	0.666667
			3	0.921466	0.714286
			4	0.916230	0.714286
			5	0.910995	0.809524
			6	0.916230	0.523810
			7	0.942408	0.666667
			8	0.931937	0.761905
			9	0.937173	0.619048
		RF	0	0.973684	0.818182
			1	0.989474	0.727273
			2	0.989529	0.809524
			3	0.994764	0.809524
			4	0.984293	0.857143
			5	0.968586	0.761905
			6	0.989529	0.619048
			7	0.979058	0.809524
			8	0.973822	0.809524
			9	0.973822	0.714286
	TSVD	KNN	0	1.000000	0.681818
			1	1.000000	0.681818
			2	1.000000	0.714286
			3	1.000000	0.857143

Continued on next page

Table 11 – Accuracy of best-performing pipelines in 10-fold stratified CV. Source: the author.

			Measure	Train accuracy	Test accuracy
Dimensionality	Projector	Classifier	Fold		
			4	1.000000	0.809524
			5	1.000000	0.761905
			6	1.000000	0.619048
			7	1.000000	0.714286
			8	1.000000	0.809524
			9	1.000000	0.619048
		MLP	0	0.715789	0.681818
			1	0.726316	0.545455
			2	0.691099	0.714286
			3	0.753927	0.809524
			4	0.732984	0.809524
			5	0.670157	0.666667
			6	0.701571	0.666667
			7	0.696335	0.714286
			8	0.696335	0.714286
			9	0.743455	0.666667
		SVM	0	0.705263	0.727273
			1	0.736842	0.590909
			2	0.670157	0.761905
			3	0.732984	0.809524
			4	0.717277	0.809524
			5	0.675393	0.714286
			6	0.701571	0.666667
			7	0.685864	0.714286
			8	0.701571	0.619048
			9	0.743455	0.714286
		RF	0	0.857895	0.681818
			1	0.868421	0.590909
			2	0.832461	0.761905
			3	0.858639	0.857143
			4	0.853403	0.857143

Continued on next page

Table 11 – Accuracy of best-performing pipelines in 10-fold stratified CV. Source: the author.

			Measure	Train accuracy	Test accuracy
Dimensionality	Projector	Classifier	Fold		
			5	0.853403	0.714286
			6	0.816754	0.666667
			7	0.842932	0.714286
			8	0.821990	0.666667
			9	0.890052	0.666667
	PCS	KNN	0	1.000000	0.818182
			1	1.000000	0.590909
			2	1.000000	0.476190
			3	1.000000	0.523810
			4	1.000000	0.666667
			5	1.000000	0.666667
			6	1.000000	0.714286
			7	1.000000	0.666667
			8	1.000000	0.761905
			9	1.000000	0.523810
		MLP	0	0.500000	0.500000
			1	0.500000	0.500000
			2	0.497382	0.523810
			3	0.497382	0.523810
			4	0.497382	0.523810
			5	0.497382	0.523810
			6	0.502618	0.476190
			7	0.502618	0.476190
			8	0.502618	0.476190
			9	0.502618	0.476190
		SVM	0	0.552632	0.500000
			1	0.542105	0.545455
			2	0.502618	0.476190
			3	0.502618	0.476190
			4	0.502618	0.476190
			5	0.502618	0.476190

Continued on next page

Table 11 – Accuracy of best-performing pipelines in 10-fold stratified CV. Source: the author.

			Measure	Train accuracy	Test accuracy
Dimensionality	Projector	Classifier	Fold		
			6	0.502618	0.476190
			7	0.502618	0.476190
			8	0.502618	0.476190
			9	0.502618	0.476190
		RF	0	1.000000	0.727273
			1	1.000000	0.590909
			2	1.000000	0.523810
			3	1.000000	0.619048
			4	1.000000	0.476190
			5	1.000000	0.619048
			6	1.000000	0.619048
			7	1.000000	0.666667
			8	1.000000	0.666667
			9	1.000000	0.523810
	KPCA	KNN	0	1.000000	0.727273
			1	1.000000	0.590909
			2	1.000000	0.809524
			3	1.000000	0.809524
			4	1.000000	0.857143
			5	1.000000	0.761905
			6	1.000000	0.809524
			7	1.000000	0.714286
			8	1.000000	0.619048
			9	1.000000	0.666667
		MLP	0	0.726316	0.772727
			1	0.773684	0.590909
			2	0.748691	0.809524
			3	0.759162	0.809524
			4	0.722513	0.904762
			5	0.748691	0.714286
			6	0.785340	0.571429

Continued on next page

Table 11 – Accuracy of best-performing pipelines in 10-fold stratified CV. Source: the author.

			Measure	Train accuracy	Test accuracy
Dimensionality	Projector	Classifier	Fold		
			7	0.743455	0.714286
			8	0.748691	0.666667
			9	0.759162	0.666667
		SVM	0	0.715789	0.727273
			1	0.742105	0.500000
			2	0.722513	0.857143
			3	0.712042	0.761905
			4	0.722513	0.809524
			5	0.722513	0.904762
			6	0.717277	0.666667
			7	0.727749	0.761905
			8	0.738220	0.666667
			9	0.748691	0.619048
		RF	0	1.000000	0.727273
			1	1.000000	0.636364
			2	1.000000	0.857143
			3	1.000000	0.761905
			4	1.000000	0.857143
			5	1.000000	0.666667
			6	1.000000	0.857143
			7	1.000000	0.714286
			8	1.000000	0.619048
			9	1.000000	0.571429
	UMAP	KNN	0	1.000000	0.636364
			1	1.000000	0.727273
			2	1.000000	0.761905
			3	1.000000	0.857143
			4	1.000000	0.761905
			5	1.000000	0.809524
			6	1.000000	0.809524
			7	1.000000	0.809524

Continued on next page

Table 11 – Accuracy of best-performing pipelines in 10-fold stratified CV. Source: the author.

			Measure	Train accuracy	Test accuracy
Dimensionality	Projector	Classifier	Fold		
			8	1.000000	0.761905
			9	1.000000	0.761905
		MLP	0	0.500000	0.500000
			1	0.500000	0.500000
			2	0.502618	0.476190
			3	0.502618	0.476190
			4	0.502618	0.476190
			5	0.502618	0.476190
			6	0.497382	0.523810
			7	0.497382	0.523810
			8	0.497382	0.523810
			9	0.497382	0.523810
		SVM	0	1.000000	0.590909
			1	1.000000	0.590909
			2	0.502618	0.476190
			3	0.502618	0.476190
			4	0.502618	0.476190
			5	0.502618	0.476190
			6	0.502618	0.476190
			7	0.502618	0.476190
			8	0.502618	0.476190
			9	0.502618	0.476190
		RF	0	1.000000	0.636364
			1	1.000000	0.590909
			2	1.000000	0.809524
			3	1.000000	0.714286
			4	1.000000	0.666667
			5	1.000000	0.809524
			6	1.000000	0.714286
			7	1.000000	0.666667
			8	1.000000	0.619048

Continued on next page

Table 11 – Accuracy of best-performing pipelines in 10-fold stratified CV. Source: the author.

			Measure	Train accuracy	Test accuracy
Dimensionality	Projector	Classifier	Fold		
			9	1.000000	0.666667
	Ivis	KNN	0	0.947368	0.727273
			1	0.968421	0.681818
			2	0.916230	0.857143
			3	0.942408	0.857143
			4	0.963351	0.952381
			5	0.942408	0.857143
			6	0.916230	0.857143
			7	0.958115	0.809524
			8	0.931937	0.809524
			9	0.979058	0.809524
		MLP	0	0.994737	0.772727
			1	0.989474	0.772727
			2	1.000000	0.857143
			3	1.000000	0.904762
			4	0.968586	0.952381
			5	0.994764	0.904762
			6	0.994764	0.761905
			7	0.984293	0.761905
			8	0.994764	0.857143
			9	0.984293	0.857143
		SVM	0	0.994737	0.772727
			1	0.984211	0.772727
			2	0.979058	0.904762
			3	0.989529	0.904762
			4	0.989529	0.952381
			5	1.000000	0.952381
			6	0.994764	0.809524
			7	1.000000	0.761905
			8	0.979058	0.904762
			9	1.000000	0.809524

Continued on next page

Table 11 – Accuracy of best-performing pipelines in 10-fold stratified CV. Source: the author.

			Measure	Train accuracy	Test accuracy
Dimensionality	Projector	Classifier	Fold		
		RF	0	1.000000	0.681818
			1	0.994737	0.636364
			2	1.000000	0.714286
			3	1.000000	0.666667
			4	1.000000	0.761905
			5	1.000000	0.761905
			6	1.000000	0.571429
			7	1.000000	0.714286
			8	1.000000	0.666667
			9	1.000000	0.666667

APPENDIX

B

ADDITIONAL *POST-HOC* TEST RESULTS OF PIPELINE TUNING SCORES

Tables 12 and 13 expose the results of *post-hoc* tests on train and validation accuracy scores. Comparisons were made between all possible pipeline pairs; however, considering that the focus of this study lies on comparing DR-encompassing pipelines against pipelines trained on the original set, only comparisons between pipelines with and without DR are displayed. Comparisons involving only one type of pipeline were omitted.

The produced tabular visualizations consist of rows that correspond to a pipeline with DR and columns that correspond to a baseline pipeline. The DR-encompassing pipelines are stratified by dimensionality, projector, and classifier, whereas the baseline ones are grouped by *post-hoc* test and p-value adjustment technique (if applicable).

By inspecting the p-values of the train accuracy scores on Table 12, one can notice that multiple comparisons involving the baseline MLP pipeline resulted in p-values substantially below the significance level. Upon inspection of these pairs of pipelines, one might conclude that this is because the scores of the baseline MLP pipeline are among the worst ones. Many pipelines attained considerably better scores to the point that it is unlikely that this discrepancy happened by chance.

To a lesser degree, comparisons involving the baseline RF pipeline also resulted in p-values below the 0.05 threshold. In this case, however, further scrutiny reveals that this is a consequence of the performance of the baseline RF pipeline, which is among the best observed. As such, the DR-encompassing pipelines that participated in these comparisons must be significantly worst-performing in labeling the train set to the point that the null hypothesis can be safely discarded. Additional noteworthy aspects of Table 12 include the fact that no comparison involving the baseline KNN and SVM pipelines, as well as the DR-encompassing ones that used PCA, resulted in p-values below significance level.

Moving on to the statistical significance results of the validation accuracy scores seen in Table 13, one might notice that the occurrences of p-values that are lower than 0.05 are more scattered, as there are also comparisons involving the baseline SVM pipeline that scored low p-values. Only comparisons involving the baseline KNN pipeline or pipelines that used TSVD for DR recurrently obtained p-values above significance level.

Table 12 – Friedman *post-hoc* tests on train accuracy scores between pipelines without and with DR. Source: the author.

		Test	Conover												Nemenyi			
		p-value adjustment technique	—				Bonferroni				Hommel				—			
		Classifier	KNN	MLP	SVM	RF	KNN	MLP	SVM	RF	KNN	MLP	SVM	RF	KNN	MLP	SVM	RF
Dimensionality	Projector	Classifier																
2	PCA	KNN	0.0004	0.0000	0.0307	0.7845	0.5802	0.0001	1.0000	1.0000	0.3107	0.0001	1.0000	1.0000	0.1716	0.0010	0.9000	0.9000
		MLP	0.4853	0.1836	0.0387	0.0001	1.0000	1.0000	1.0000	0.1120	1.0000	1.0000	1.0000	0.0725	0.9000	0.9000	0.9000	0.0412
		SVM	0.3881	0.0040	0.6095	0.0166	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	0.9000	0.6416	0.9000	0.9000
		RF	0.4544	0.0057	0.5314	0.0121	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	0.9000	0.7244	0.9000	0.9000
	TSVD	KNN	0.0004	0.0000	0.0307	0.7845	0.5802	0.0001	1.0000	1.0000	0.3107	0.0001	1.0000	1.0000	0.1716	0.0010	0.9000	0.9000
		MLP	0.1743	0.5035	0.0065	0.0000	1.0000	1.0000	1.0000	0.0063	1.0000	1.0000	1.0000	0.0049	0.9000	0.9000	0.7554	0.0024
		SVM	0.3038	0.0023	0.7298	0.0256	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	0.9000	0.5225	0.9000	0.9000
		RF	0.0004	0.0000	0.0307	0.7845	0.5802	0.0001	1.0000	1.0000	0.3107	0.0001	1.0000	1.0000	0.1716	0.0010	0.9000	0.9000
	PCS	KNN	0.3388	0.2840	0.0201	0.0000	1.0000	1.0000	1.0000	0.0381	1.0000	1.0000	1.0000	0.0262	0.9000	0.9000	0.9000	0.0146
		MLP	0.1589	0.5361	0.0056	0.0000	1.0000	1.0000	1.0000	0.0050	1.0000	1.0000	1.0000	0.0039	0.9000	0.9000	0.7192	0.0019
		SVM	0.5314	0.1610	0.0459	0.0001	1.0000	1.0000	1.0000	0.1497	1.0000	1.0000	1.0000	0.0943	0.9000	0.9000	0.9000	0.0536
		RF	0.0004	0.0000	0.0307	0.7845	0.5802	0.0001	1.0000	1.0000	0.3107	0.0001	1.0000	1.0000	0.1716	0.0010	0.9000	0.9000
	KPCA	KNN	0.8403	0.0261	0.2412	0.0023	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	0.9000	0.9000	0.9000	0.5173
		MLP	0.4330	0.2136	0.0313	0.0001	1.0000	1.0000	1.0000	0.0787	1.0000	1.0000	1.0000	0.0520	0.9000	0.9000	0.9000	0.0295
		SVM	0.4898	0.1812	0.0394	0.0001	1.0000	1.0000	1.0000	0.1153	1.0000	1.0000	1.0000	0.0745	0.9000	0.9000	0.9000	0.0423
		RF	0.0007	0.0000	0.0429	0.8913	0.9534	0.0001	1.0000	1.0000	0.4745	0.0001	1.0000	1.0000	0.2494	0.0010	0.9000	0.9000
	UMAP	KNN	0.0004	0.0000	0.0307	0.7845	0.5802	0.0001	1.0000	1.0000	0.3107	0.0001	1.0000	1.0000	0.1716	0.0010	0.9000	0.9000
		MLP	0.0261	0.8403	0.0003	0.0000	1.0000	1.0000	0.4560	0.0001	1.0000	1.0000	0.2531	0.0001	0.9000	0.9000	0.1401	0.0010
		SVM	0.6712	0.1091	0.0725	0.0002	1.0000	1.0000	1.0000	0.3289	1.0000	1.0000	1.0000	0.1925	0.9000	0.9000	0.9000	0.1076
		RF	0.0004	0.0000	0.0307	0.7845	0.5802	0.0001	1.0000	1.0000	0.3107	0.0001	1.0000	1.0000	0.1716	0.0010	0.9000	0.9000
	Ivis	KNN	0.0004	0.0000	0.0307	0.7845	0.5802	0.0001	1.0000	1.0000	0.3107	0.0001	1.0000	1.0000	0.1716	0.0010	0.9000	0.9000
		MLP	0.0429	1.0000	0.0007	0.0000	1.0000	1.0000	0.9534	0.0002	1.0000	1.0000	0.4745	0.0002	0.9000	0.9000	0.2494	0.0010
		SVM	0.0067	0.0000	0.1789	0.5845	1.0000	0.0036	1.0000	1.0000	1.0000	0.0028	1.0000	1.0000	0.7658	0.0013	0.9000	0.9000
		RF	0.0004	0.0000	0.0307	0.7845	0.5802	0.0001	1.0000	1.0000	0.3107	0.0001	1.0000	1.0000	0.1716	0.0010	0.9000	0.9000
3	PCA	KNN	0.6045	0.0111	0.3920	0.0062	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	0.9000	0.8900	0.9000	0.7451
		MLP	0.9255	0.0342	0.2007	0.0016	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	0.8734	0.9000	0.9000	0.9000	0.4334
		SVM	0.1387	0.0005	0.9141	0.0749	1.0000	0.6450	1.0000	1.0000	1.0000	0.3400	1.0000	1.0000	0.9000	0.1854	0.9000	0.9000
		RF	0.0660	0.0001	0.6400	0.1547	1.0000	0.1631	1.0000	1.0000	1.0000	0.1021	1.0000	1.0000	0.9000	0.0582	0.9000	0.9000
	TSVD	KNN	0.0004	0.0000	0.0307	0.7845	0.5802	0.0001	1.0000	1.0000	0.3107	0.0001	1.0000	1.0000	0.1716	0.0010	0.9000	0.9000
		MLP	0.5995	0.1331	0.0580	0.0002	1.0000	1.0000	1.0000	0.2228	1.0000	1.0000	1.0000	0.1354	0.9000	0.9000	0.9000	0.0775
		SVM	0.4808	0.1859	0.0380	0.0001	1.0000	1.0000	1.0000	0.1088	1.0000	1.0000	1.0000	0.0706	0.9000	0.9000	0.9000	0.0401
		RF	0.2383	0.0014	0.8460	0.0374	1.0000	1.0000	1.0000	1.0000	1.0000	0.8015	1.0000	1.0000	0.9000	0.4016	0.9000	0.9000
	PCS	KNN	0.0004	0.0000	0.0307	0.7845	0.5802	0.0001	1.0000	1.0000	0.3107	0.0001	1.0000	1.0000	0.1716	0.0010	0.9000	0.9000
		MLP	0.0261	0.8403	0.0003	0.0000	1.0000	1.0000	0.4560	0.0001	1.0000	1.0000	0.2531	0.0001	0.9000	0.9000	0.1401	0.0010
		SVM	0.0525	0.9312	0.0010	0.0000	1.0000	1.0000	1.0000	0.0004	1.0000	1.0000	0.6109	0.0003	0.9000	0.9000	0.3102	0.0010
		RF	0.0004	0.0000	0.0307	0.7845	0.5802	0.0001	1.0000	1.0000	0.3107	0.0001	1.0000	1.0000	0.1716	0.0010	0.9000	0.9000
	KPCA	KNN	0.0004	0.0000	0.0307	0.7845	0.5802	0.0001	1.0000	1.0000	0.3107	0.0001	1.0000	1.0000	0.1716	0.0010	0.9000	0.9000
		MLP	0.7680	0.0205	0.2808	0.0031	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	0.9000	0.9000	0.9000	0.5846
		SVM	0.8403	0.0681	0.1156	0.0006	1.0000	1.0000	1.0000	0.7551	1.0000	1.0000	1.0000	0.3889	0.9000	0.9000	0.9000	0.2099

Continued on next page

Table 12 – Friedman *post-hoc* tests on train accuracy scores between pipelines without and with DR. Source: the author.

		Test	Conover												Nemenyi			
		p-value adjustment technique	—				Bonferroni				Hommel				—			
		Classifier	KNN	MLP	SVM	RF	KNN	MLP	SVM	RF	KNN	MLP	SVM	RF	KNN	MLP	SVM	RF
Dimensionality	Projector	Classifier																
		RF	0.0004	0.0000	0.0307	0.7845	0.5802	0.0001	1.0000	1.0000	0.3107	0.0001	1.0000	1.0000	0.1716	0.0010	0.9000	0.9000
	UMAP	KNN	0.0004	0.0000	0.0307	0.7845	0.5802	0.0001	1.0000	1.0000	0.3107	0.0001	1.0000	1.0000	0.1716	0.0010	0.9000	0.9000
		MLP	0.0261	0.8403	0.0003	0.0000	1.0000	1.0000	0.4560	0.0001	1.0000	1.0000	0.2531	0.0001	0.9000	0.9000	0.1401	0.0010
		SVM	0.3881	0.2441	0.0256	0.0000	1.0000	1.0000	1.0000	0.0566	1.0000	1.0000	1.0000	0.0382	0.9000	0.9000	0.9000	0.0215
		RF	0.0004	0.0000	0.0307	0.7845	0.5802	0.0001	1.0000	1.0000	0.3107	0.0001	1.0000	1.0000	0.1716	0.0010	0.9000	0.9000
	Ivis	KNN	0.1091	0.0003	0.8179	0.0970	1.0000	0.4093	1.0000	1.0000	1.0000	0.2318	1.0000	1.0000	0.9000	0.1276	0.9000	0.9000
		MLP	0.0225	0.0000	0.3610	0.3280	1.0000	0.0254	1.0000	1.0000	1.0000	0.0179	1.0000	1.0000	0.9000	0.0098	0.9000	0.9000
		SVM	0.0172	0.0000	0.3106	0.3802	1.0000	0.0163	1.0000	1.0000	1.0000	0.0119	1.0000	1.0000	0.9000	0.0063	0.9000	0.9000
		RF	0.0007	0.0000	0.0429	0.8913	0.9534	0.0001	1.0000	1.0000	0.4745	0.0001	1.0000	1.0000	0.2494	0.0010	0.9000	0.9000

Table 13 – Friedman *post-hoc* tests on validation accuracy scores between pipelines without and with DR. Source: the author.

		Test	Conover												Nemenyi			
		p-value adjustment technique	—				Bonferroni				Hommel				—			
		Classifier	KNN	MLP	SVM	RF	KNN	MLP	SVM	RF	KNN	MLP	SVM	RF	KNN	MLP	SVM	RF
Dimensionality	Projector	Classifier																
2	PCA	KNN	0.3099	0.0010	0.1253	0.0154	1.0000	1.0000	1.0000	1.0000	1.0000	0.7179	1.0000	1.0000	0.9000	0.2792	0.9000	0.9000
		MLP	0.5134	0.0033	0.0583	0.0054	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	0.9000	0.5691	0.9000	0.6881
		SVM	0.6751	0.0067	0.0335	0.0026	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	0.9990	0.9000	0.7399	0.9000	0.5173
		RF	0.0565	0.0000	0.5226	0.1253	1.0000	0.0398	1.0000	1.0000	1.0000	0.0337	1.0000	1.0000	0.9000	0.0121	0.9000	0.9000
	TSVD	KNN	0.7331	0.0085	0.0275	0.0020	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	0.9967	0.9000	0.7969	0.9000	0.4577
		MLP	0.8146	0.0391	0.0055	0.0003	1.0000	1.0000	1.0000	0.3428	1.0000	1.0000	1.0000	0.2384	0.9000	0.9000	0.6933	0.0949
		SVM	0.9264	0.0275	0.0085	0.0004	1.0000	1.0000	1.0000	0.5842	1.0000	1.0000	1.0000	0.3732	0.9000	0.9000	0.7969	0.1503
		RF	0.6699	0.0066	0.0341	0.0027	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	0.9993	0.9000	0.7347	0.9000	0.5225
	PCS	KNN	0.8926	0.0151	0.0160	0.0010	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	0.7318	0.9000	0.9000	0.9000	0.2843
		MLP	0.6855	0.0583	0.0033	0.0001	1.0000	1.0000	1.0000	0.1767	1.0000	1.0000	1.0000	0.1322	0.9000	0.9000	0.5691	0.0522
		SVM	0.8479	0.0129	0.0187	0.0012	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	0.8379	0.9000	0.9000	0.9000	0.3259
		RF	0.2330	0.2680	0.0002	0.0000	1.0000	1.0000	0.2682	0.0060	1.0000	1.0000	0.1911	0.0055	0.9000	0.9000	0.0775	0.0017
	KPCA	KNN	0.1027	0.0001	0.3596	0.0705	1.0000	0.1257	1.0000	1.0000	1.0000	0.0977	1.0000	1.0000	0.9000	0.0380	0.9000	0.9000
		MLP	0.5991	0.0049	0.0433	0.0037	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	0.9000	0.6623	0.9000	0.5950
		SVM	0.4688	0.0026	0.0684	0.0067	1.0000	1.0000	1.0000	1.0000	1.0000	0.9990	1.0000	1.0000	0.9000	0.5173	0.9000	0.7399
		RF	0.2966	0.0009	0.1325	0.0167	1.0000	1.0000	1.0000	1.0000	1.0000	0.6636	1.0000	1.0000	0.9000	0.2591	0.9000	0.9000
	UMAP	KNN	0.0225	0.0000	0.7926	0.2472	1.0000	0.0075	1.0000	1.0000	1.0000	0.0069	1.0000	1.0000	0.9000	0.0021	0.9000	0.9000

Continued on next page

Table 13 – Friedman *post-hoc* tests on validation accuracy scores between pipelines without and with DR. Source: the author.

		Test															
										Conover				Nemenyi			
		p-value adjustment technique	—				Bonferroni				Hommel				—			
		Classifier	KNN	MLP	SVM	RF	KNN	MLP	SVM	RF	KNN	MLP	SVM	RF	KNN	MLP	SVM	RF
Dimensionality	Projector	Classifier																
		MLP	0.0456	0.7654	0.0000	0.0000	1.0000	1.0000	0.0088	0.0001	1.0000	1.0000	0.0081	0.0001	0.9000	0.9000	0.0025	0.0010
		SVM	0.8870	0.0312	0.0073	0.0004	1.0000	1.0000	1.0000	0.4858	1.0000	1.0000	1.0000	0.3202	0.9000	0.9000	0.7606	0.1276
		RF	0.0177	0.0000	0.8646	0.2868	1.0000	0.0049	1.0000	1.0000	1.0000	0.0046	1.0000	1.0000	0.9000	0.0013	0.9000	0.9000
	Ivis	KNN	0.0003	0.0000	0.2901	0.8702	0.4485	0.0000	1.0000	1.0000	0.2994	0.0000	1.0000	1.0000	0.1216	0.0010	0.9000	0.9000
		MLP	0.0217	1.0000	0.0000	0.0000	1.0000	1.0000	0.0022	0.0000	1.0000	1.0000	0.0021	0.0000	0.9000	0.9000	0.0010	0.0010
		SVM	0.0006	0.0000	0.3785	0.9887	0.8615	0.0000	1.0000	1.0000	0.5184	0.0000	1.0000	1.0000	0.2056	0.0010	0.9000	0.9000
		RF	0.0010	0.0000	0.4473	0.8926	1.0000	0.0000	1.0000	1.0000	0.7318	0.0000	1.0000	1.0000	0.2843	0.0010	0.9000	0.9000
3	PCA	KNN	0.1343	0.0002	0.2933	0.0520	1.0000	0.2150	1.0000	1.0000	1.0000	0.1569	1.0000	1.0000	0.9000	0.0630	0.9000	0.9000
		MLP	0.0705	0.0000	0.4601	0.1027	1.0000	0.0606	1.0000	1.0000	1.0000	0.0499	1.0000	1.0000	0.9000	0.0186	0.9000	0.9000
		SVM	0.3901	0.0017	0.0913	0.0100	1.0000	1.0000	1.0000	1.0000	1.0000	0.9824	1.0000	1.0000	0.9000	0.4080	0.9000	0.8383
		RF	0.0106	0.0000	0.9887	0.3785	1.0000	0.0020	1.0000	1.0000	1.0000	0.0019	1.0000	1.0000	0.8538	0.0010	0.9000	0.9000
	TSVD	KNN	0.1685	0.0003	0.2414	0.0391	1.0000	0.3428	1.0000	1.0000	1.0000	0.2384	1.0000	1.0000	0.9000	0.0949	0.9000	0.9000
		MLP	0.4222	0.0020	0.0810	0.0085	1.0000	1.0000	1.0000	1.0000	1.0000	0.9967	1.0000	1.0000	0.9000	0.4577	0.9000	0.7969
		SVM	0.2901	0.0008	0.1362	0.0173	1.0000	1.0000	1.0000	1.0000	1.0000	0.6383	1.0000	1.0000	0.9000	0.2494	0.9000	0.9000
		RF	0.2302	0.0005	0.1775	0.0252	1.0000	0.6657	1.0000	1.0000	1.0000	0.4172	1.0000	1.0000	0.9000	0.1680	0.9000	0.9000
	PCS	KNN	0.8590	0.0135	0.0180	0.0012	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	0.8127	0.9000	0.9000	0.9000	0.3154
		MLP	0.0456	0.7654	0.0000	0.0000	1.0000	1.0000	0.0088	0.0001	1.0000	1.0000	0.0081	0.0001	0.9000	0.9000	0.0025	0.0010
		SVM	0.0285	0.9151	0.0000	0.0000	1.0000	1.0000	0.0036	0.0000	1.0000	1.0000	0.0034	0.0000	0.9000	0.9000	0.0010	0.0010
		RF	0.5180	0.0983	0.0015	0.0000	1.0000	1.0000	1.0000	0.0662	1.0000	1.0000	0.9334	0.0542	0.9000	0.9000	0.3757	0.0203
	KPCA	KNN	0.0969	0.0001	0.3747	0.0750	1.0000	0.1121	1.0000	1.0000	1.0000	0.0880	1.0000	1.0000	0.9000	0.0340	0.9000	0.9000
		MLP	0.1867	0.0003	0.2194	0.0341	1.0000	0.4252	1.0000	1.0000	1.0000	0.2860	1.0000	1.0000	0.9000	0.1159	0.9000	0.9000
		SVM	0.1797	0.0003	0.2275	0.0359	1.0000	0.3924	1.0000	1.0000	1.0000	0.2672	1.0000	1.0000	0.9000	0.1076	0.9000	0.9000
		RF	0.1685	0.0003	0.2414	0.0391	1.0000	0.3428	1.0000	1.0000	1.0000	0.2384	1.0000	1.0000	0.9000	0.0949	0.9000	0.9000
	UMAP	KNN	0.0183	0.0000	0.8534	0.2805	1.0000	0.0052	1.0000	1.0000	1.0000	0.0049	1.0000	1.0000	0.9000	0.0014	0.9000	0.9000
		MLP	0.0405	0.8036	0.0000	0.0000	1.0000	1.0000	0.0070	0.0001	1.0000	1.0000	0.0065	0.0001	0.9000	0.9000	0.0020	0.0010
		SVM	0.0433	0.7817	0.0000	0.0000	1.0000	1.0000	0.0080	0.0001	1.0000	1.0000	0.0074	0.0001	0.9000	0.9000	0.0023	0.0010
		RF	0.5602	0.0041	0.0495	0.0044	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	0.9000	0.6209	0.9000	0.6364
	Ivis	KNN	0.0007	0.0000	0.3901	0.9717	0.9300	0.0000	1.0000	1.0000	0.5532	0.0000	1.0000	1.0000	0.2185	0.0010	0.9000	0.9000
		MLP	0.0002	0.0000	0.2330	0.7654	0.2682	0.0000	1.0000	1.0000	0.1911	0.0000	1.0000	1.0000	0.0775	0.0010	0.9000	0.9000
		SVM	0.0001	0.0000	0.1752	0.6442	0.1409	0.0000	1.0000	1.0000	0.1084	0.0000	1.0000	1.0000	0.0423	0.0010	0.9000	0.9000
		RF	0.6493	0.0060	0.0365	0.0029	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	0.9000	0.7140	0.9000	0.5432

APPENDIX

C

ADDITIONAL PROJECTION FIGURES

Complementing the visualizations exposed in Section 4.2.1, this appendix contains representations of the merged projections produced for the train and test sets.

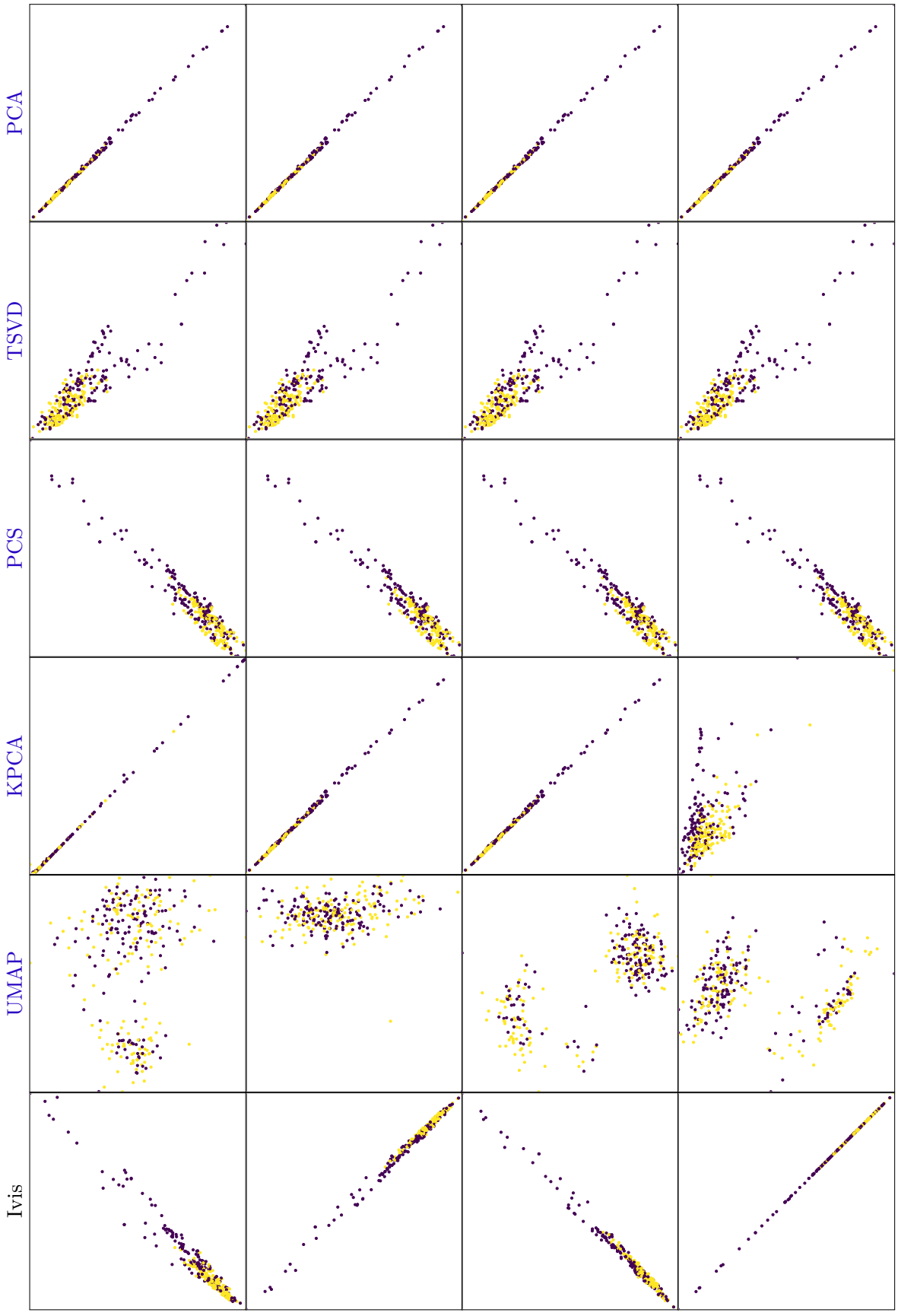
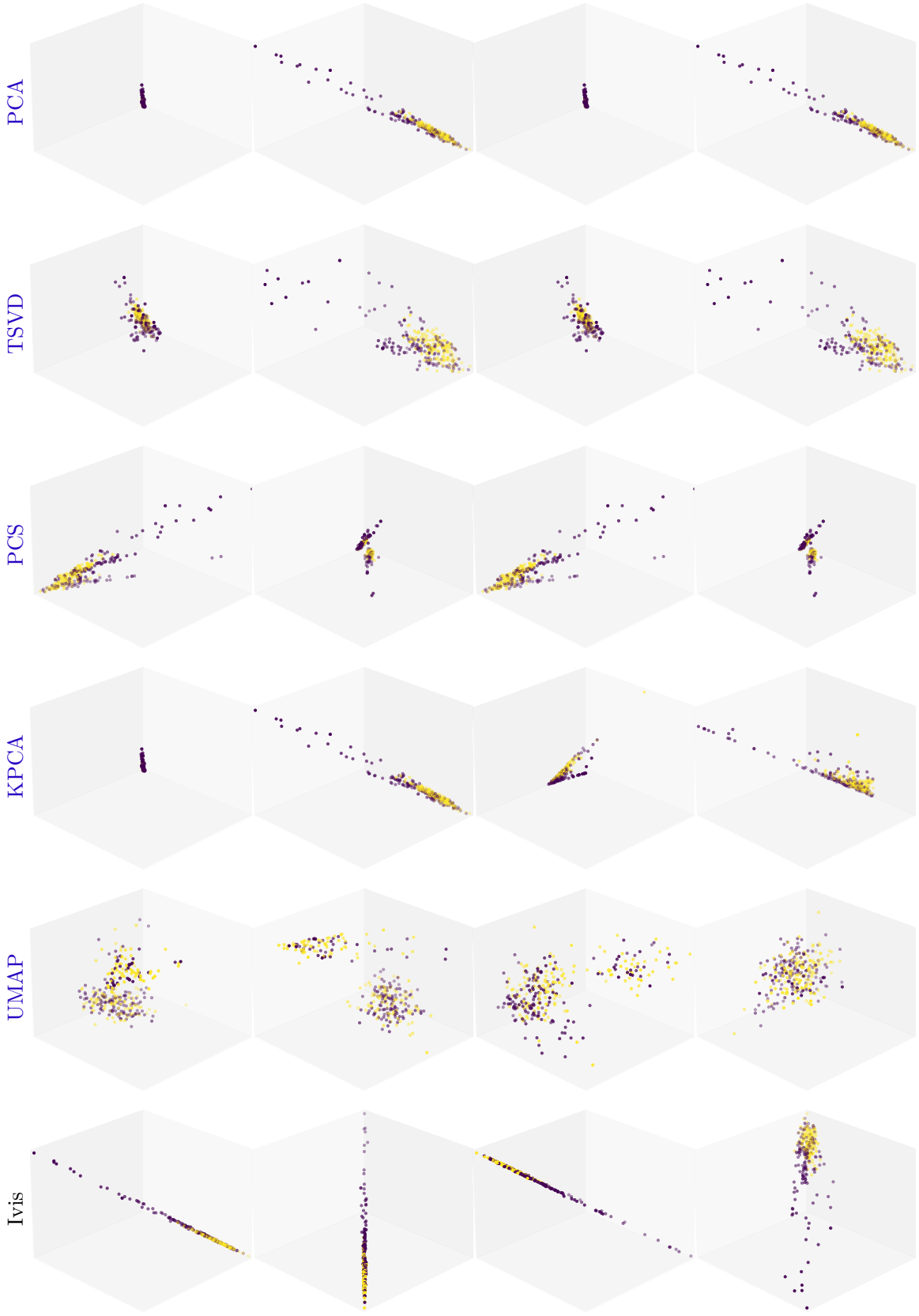
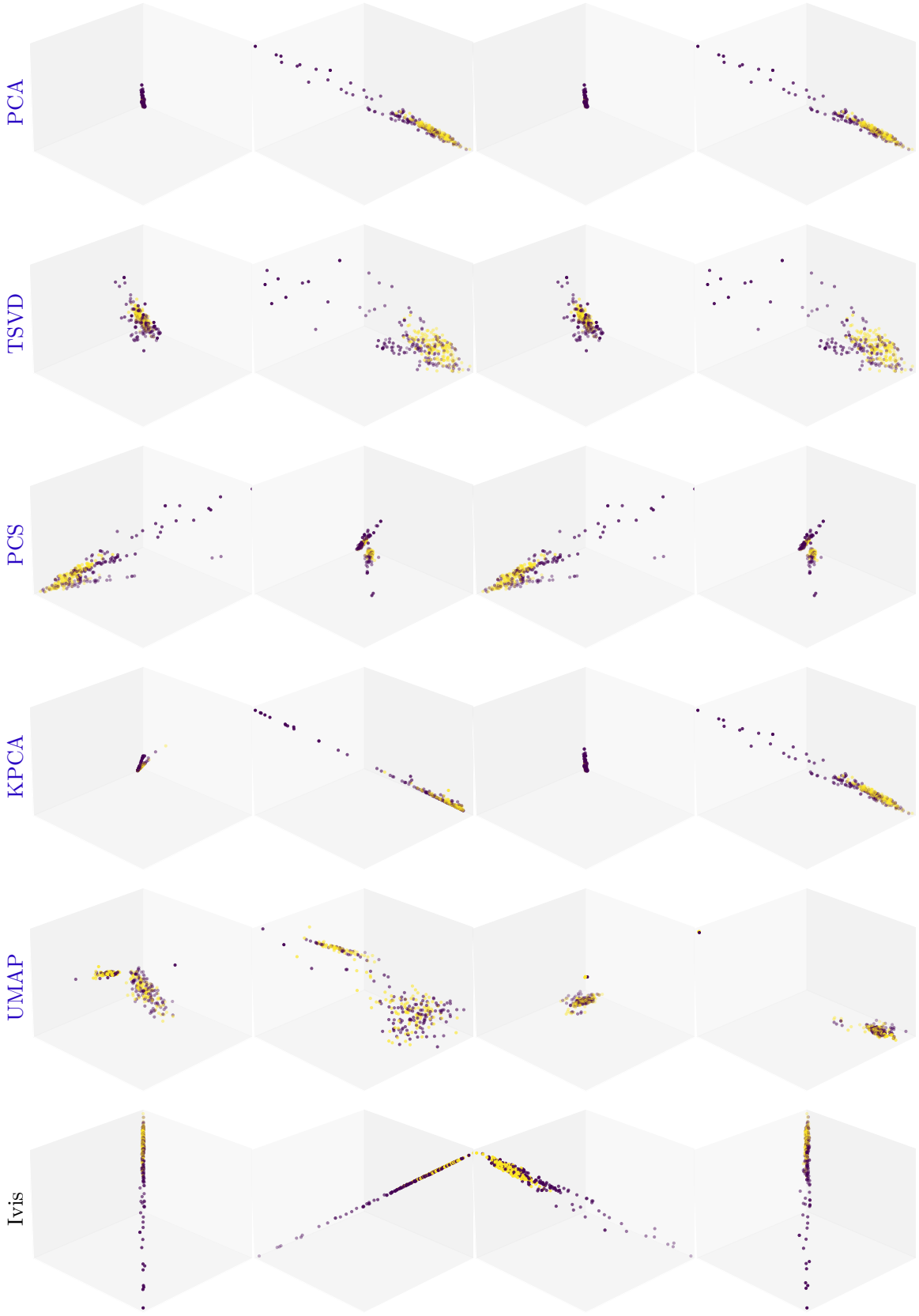


Figure 19 – Two-dimensional projections of the entire set for all classifiers, with HIA (–) samples represented in purple and HIA (+) in yellow. Source: the author.



(a) KNN (b) MLP

Figure 20 – Two perspectives of three-dimensional projections of the entire set for KNN and MLP, with HIA (–) samples represented in purple and HIA (+) in yellow. Source: the author.



(a) SVM (b) RF

Figure 21 – Two perspectives of three-dimensional projections of the entire set for SVM and RF, with HIA (–) samples represented in purple and HIA (+) in yellow. Source: the author.