POLITECHNIKA GDAŃSKA

Imię i nazwisko autora rozprawy: Karol Baran
Dyscyplina naukowa: nauki chemiczne

ROZPRAWA DOKTORSKA

Tytuł rozprawy w języku polskim: Uczenie maszynowe na kilku przykładach w modelowaniu układów multimolekularnych

Tytuł rozprawy w języku angielskim: Few-shot machine learning for multimolecular systems modeling

Promotor *podpis*
prof. dr hab. inż. Adam Kloskowski

Gdańsk, rok 2026

POLITECHNIKA GDAŃSKA

OŚWIADCZENIE

Autor rozprawy doktorskiej: Karol Baran

Ja, niżej podpisany(a), oświadczam, iż jestem świadomy(a), że zgodnie z przepisem art. 27 ust. 1 i 2 ustawy z dnia 4 lutego 1994 r. o prawie autorskim i prawach pokrewnych (t.j. Dz.U. z 2021 poz. 1062), uczelnia może korzystać z mojej rozprawy doktorskiej zatytułowanej: „Few-shot machine learning for multimolecular systems modeling" do prowadzenia badań naukowych lub w celach dydaktycznych.[1]

Świadomy(a) odpowiedzialności karnej z tytułu naruszenia przepisów ustawy z dnia 4 lutego 1994 r. o prawie autorskim i prawach pokrewnych i konsekwencji dyscyplinarnych określonych w ustawie Prawo o szkolnictwie wyższym i nauce (Dz.U.2021.478 t.j.), a także odpowiedzialności cywilno-prawnej oświadczam, że przedkładana rozprawa doktorska została napisana przeze mnie samodzielnie.

Oświadczam, że treść rozprawy opracowana została na podstawie wyników badań prowadzonych pod kierunkiem i w ścisłej współpracy z promotorem Adamem Kloskowskim.

Niniejsza rozprawa doktorska nie była wcześniej podstawą żadnej innej urzędowej procedury związanej z nadaniem stopnia doktora.

Wszystkie informacje umieszczone w ww. rozprawie uzyskane ze źródeł pisanych i elektronicznych, zostały udokumentowane w wykazie literatury odpowiednimi odnośnikami, zgodnie z przepisem art. 34 ustawy o prawie autorskim i prawach pokrewnych.

Potwierdzam zgodność niniejszej wersji pracy doktorskiej z załączoną wersją elektroniczną.

Gdańsk, dnia ....

podpis doktoranta

Ja, niżej podpisany(a), wyrażam zgodę/~~nie wyrażam zgody~~* na umieszczenie ww. rozprawy doktorskiej w wersji elektronicznej w otwartym, cyfrowym repozytorium instytucjonalnym Politechniki Gdańskiej.

Gdańsk, dnia ....

podpis doktoranta

**niepotrzebne usunąć*

[1] Art. 27. 1. Instytucje oświatowe oraz podmioty, o których mowa w art. 7 ust. 1 pkt 1, 2 i 4–8 ustawy z dnia 20 lipca 2018 r. – Prawo o szkolnictwie wyższym i nauce, mogą na potrzeby zilustrowania treści przekazywanych w celach dydaktycznych lub w celu prowadzenia działalności naukowej korzystać z rozpowszechnionych utworów w oryginale i w tłumaczeniu oraz zwielokrotniać w tym celu rozpowszechnione drobne utwory lub fragmenty większych utworów.
2. W przypadku publicznego udostępniania utworów w taki sposób, aby każdy mógł mieć do nich dostęp w miejscu i czasie przez siebie wybranym korzystanie, o którym mowa w ust. 1, jest dozwolone wyłącznie dla ograniczonego kręgu osób uczących się, nauczających lub prowadzących badania naukowe, zidentyfikowanych przez podmioty wymienione w ust. 1.

GDAŃSK UNIVERSITY OF TECHNOLOGY

OPIS ROZPRAWY DOKTORSKIEJ

Autor rozprawy doktorskiej: Karol Baran

Tytuł rozprawy doktorskiej w języku polskim: Uczenie maszynowe na kilku przykładach w modelowaniu układów multimolekularnych

Tytuł rozprawy w języku angielskim: Few-shot machine learning for multimolecular systems modeling

Język rozprawy doktorskiej: angielski

Promotor rozprawy doktorskiej: Adam Kloskowski

Słowa kluczowe rozprawy doktorskiej w języku polskim: sztuczna inteligencja, przewidywanie właściwości molekularnych, algorytmy, chemoinformatyka

Słowa kluczowe rozprawy doktorskiej w języku angielskim: artificial intelligence, molecular property prediction, algorithms, chemoinformatics

Streszczenie rozprawy w języku polskim:

Zastosowanie sztucznej inteligencji w domenach naukowych często wymaga implementacji metodologii uczenia z niewielką liczbą przykładów (few-shot learning). Niniejsze badanie analizuje te techniki w kontekście modelowania molekularnego, wykorzystując zróżnicowane zbiory danych obejmujące właściwości fizykochemiczne, toksyczność oraz sorpcję z cieczami jonowymi. Początkowo zastosowano klasyczne podejścia uczenia maszynowego w celu wykazania ich skuteczności przy ograniczonych zbiorach danych, zwłaszcza po wzbogaceniu ich o deskryptory specyficzne dla danej dziedziny. Następna analiza metod uczenia głębokiego pozwoliła zidentyfikować optymalne strategie modelowania systemów bimolekularnych, jednocześnie wskazując na inherentne ograniczenia związane z jakością i objętością danych. Niemniej jednak, badanie wykazało, że techniki meta-uczenia przewyższyły metody klasyczne w reżimach o niskiej dostępności danych. W konsekwencji, praca proponuje ukierunkowane adaptacje domenowe: Attentive Model-Agnostic Meta-Learning do przewidywania toksyczności oraz Task Similarity Aware Reptile do modelowania sorpcji.

Streszczenie rozprawy w języku angielskim:

The application of Artificial Intelligence within scientific domains frequently necessitates the deployment of few-shot learning methodologies. This investigation examines these techniques specifically within the framework of molecular modeling, utilizing a variety of datasets encompassing physicochemical properties, toxicity, and sorption with ionic liquids. Initially, classical machine learning approaches were employed to demonstrate their efficacy with limited data, particularly when augmented by domain-specific descriptors. Subsequent analysis of deep learning methods identified optimal strategies for modeling bimolecular systems, while simultaneously highlighting inherent constraints related to data quality and volume. Nevertheless, the study observed that meta-learning techniques surpassed classical methods within low-data regimes. Consequently, the research proposes targeted domain adaptations: Attentive Model-Agnostic Meta-Learning for toxicity prediction and Task Similarity Aware Reptile for sorption modeling.

Writing these words, I feel grateful.

To the one who supported me since helping me with my telescope many years ago - for starting my journey to find what is interesting.

To my mentors - for teaching me that science and engineering are about finding what is meaningful.

To my coworkers - for discussions that helped me see beyond the obvious.

To my family - who anchored me in gentleness before I learned to build anything else.

To my friends - the one who knew which smiles I was faking and heard me while we were just silently walking around; the one who was there for me since the alchemy of the high school competition; the one whose waves always resonated with mine's; and all the ones with whom I shared the gift of slow conversation.

Contents

Acronyms

AE	autoencoder
AI	Artificial Intelligence
ARKA	Arithmetic Residuals in K-Groups Analysis
AUC	area under the receiver operating characteristic curve
BERT	Bidirectional encoder representations from transformers
CNN	Convolutional Neural Networks
CS	Computer Science
CV	cross-validation
CVAE	conditional variational autoencoder
DL	deep learning
FFV	Fractional Free Volume
GA	Genetic Algorithm
GB	Gradient Boosting
GNN	Graph Neural Networks
GRU	gated recurrent unit
IDAC	infinite dilution activity coefficients
IFM	Independent Feature Mapping
ILs	ionic liquids
IoU	Intersection over Union
kNN	k-nearest neighbors algorithm
LIME	Local Interpretable Model-agnostic Explanations
MAE	mean averaged error
MAML	Model-Agnostic Meta-Learning
MD	molecular descriptors
MF	molecular fingerprints
ML	machine learning
MLP	Multilayer Perceptron
MLR	Multiple Linear Regression
MPP	molecular property prediction
NN	Neural Networks
RF	Random Forest
RMSE	root mean squared error
RNN	Recurrent Neural Networks
SELFIES	SELF-referencIng Embedded Strings
SHAP	Shapley Additive Explanations
SMILES	Simplified Molecular Input Line Entry Specification
SVM	Supported Vector Machine

t-SNE	t-distributed Stochastic Neighbor Embedding
VAE	variational autoencoder
ViT	Vision Transformers

Chapter 1

Introduction

1.1 Artificial intelligence

Artificial Intelligence (AI) is revolutionizing scientific research by facilitating a discovery process that is significantly more time- and cost-efficient. Beyond enhancing accuracy, these technologies are anticipated to automate complex workflows [1]. Historically, the relationship between computing and intelligence has been a subject of rigorous debate. In 1950, Alan Turing defined a seminal test grounded in behavioral distinguishability, followed by brainstorming sessions at the 1956 Dartmouth Conference, which explored the conceptualization of thinking machines [2]. Early AI concepts were primarily grounded in rule-based systems and automata theory [3], [4]. Subsequently, the community's interest shifted toward statistical learning and probabilistic methods, including works by Hinton and Hopfield, who were awarded the Nobel Prize in Physics in 2024 [5], [6]. Among AI techniques, machine learning (ML) and deep learning (DL) are distinguished as methodologies that rely on data to construct predictive models [7]. ML techniques aim to learn from data to generate predictions. The objective is to transition from traditional explicit programming to a paradigm wherein computational systems learn patterns and insights from data autonomously. This shift is particularly critical when facing data complexities or volumes that exceed manual human analysis capabilities [8].

Data is typically structured within databases, where entries containing related items—often organized in a tabular format—are designated as records. In this context, a record corresponds to a row, while columns represent attributes or features associated with the record [9]. Insights can be derived from such data by employing ML algorithms. Within ML, tasks are primarily categorized into two distinct objectives: supervised and unsupervised learning. Unsupervised ML involves identifying relationships within data without relying on a teacher or ground truth. Conversely, supervised ML necessitates a specific target variable column, functioning as the output. This target variable may encompass predefined discrete categories (classification) or continuous numeric values (regression) [10].

Various algorithms are utilized in the field of AI. Historically, AI relied on if-then rules built using the domain knowledge of experts [11]. However, rules could also be learned from datasets. The exploration of the solution space was facilitated by algorithms such as A* [12], which addresses problems expressible as graph structures, or Genetic Algorithm (GA) [13], which approximates solutions by imitating natural selection. Another

algorithm, namely k-nearest neighbors algorithm (kNN), estimates the output based on the similarity of the query record to training samples [14].

Nowadays, Neural Networks (NN) constitute a class of algorithms widely used in many applications [15]. They aim to mimic how neurons in the human brain process signals. NNs can approximate a wide range of known mathematical functions due to their combination of linear layers and nonlinear activation functions [16]. DL methods further add feature extraction layers that allow operations on raw data (such as text or images) without requiring feature extraction based on domain knowledge. While traditional ML utilizes a two-stage pipeline (feature extraction and classifier training), DL employs end-to-end learning [17]. Learning features requires significant datasets, however [18]. Deep neural networks are often pre-trained on large datasets and fine-tuned for the task of interest [19].

Among DL methods, Convolutional Neural Networks (CNN) utilize convolution [20], while transformers and Vision Transformers (ViT) use an attention mechanism that allows NN to focus on specific parts of the input when predicting the output [21]. Sequences are often modeled using Recurrent Neural Networks (RNN) [22]. Graph Neural Networks (GNN) are specifically designed to work with graph data [23]. NN can be utilized in both supervised and unsupervised settings.

Besides NN, decision tree algorithms are often praised for their competitive accuracy [24]. Among these, ensembles like Random Forest (RF) [25] or Gradient Boosting (GB) [26] are especially common choices in modeling studies. Linear algorithms, such as Multiple Linear Regression (MLR) or Supported Vector Machine (SVM), also retain their relevance in many practical applications [27]. A visual overview of AI, ML, and DL methods is provided in Figure 1.1. A comparison between traditional ML and DL methods is shown in Figure 1.2.

The training process for the listed ML algorithms seeks to construct a model capable of predicting a target variable from input features. Optimization is commonly achieved via backpropagation to update weights, as in NN [28], or by identifying optimal data partitions, as in decision trees [29]. This procedure is shown visually in Figure 1.3. A loss function quantifies the discrepancy between the observed labels and the model's predicted outputs. Models are trained on a training set and evaluated on a disjoint test set to ensure robustness [30]. In contrast, unsupervised modeling approaches, including K-Means, DBScan, or hierarchical clustering, focus on identifying latent patterns and structures within the data [31].

Models are expected to generalize effectively to both training data and unseen test sets to ensure reliable inference. Performance is evaluated using metrics such as root mean squared error (RMSE) and mean averaged error (MAE) for regression tasks, or area under the receiver operating characteristic curve (AUC) and the F1 score for classification [32]. Generalization to novel data is essential for operational environments. Underfitting is indicated when performance is suboptimal on both training and testing datasets. Conversely, overfitting occurs when a model lacks generalization capability; it performs well on the training set but poorly on a test set [30]. Validation sets are frequently employed to detect overfitting, typically by monitoring whether the validation loss increases during training. Proper training requires an appropriate number of epochs to adjust weights; however, hyperparameters cannot be learned during training. Instead, their selection often utilizes a cross-validation (cross-validation (CV)) protocol, which splits the data into

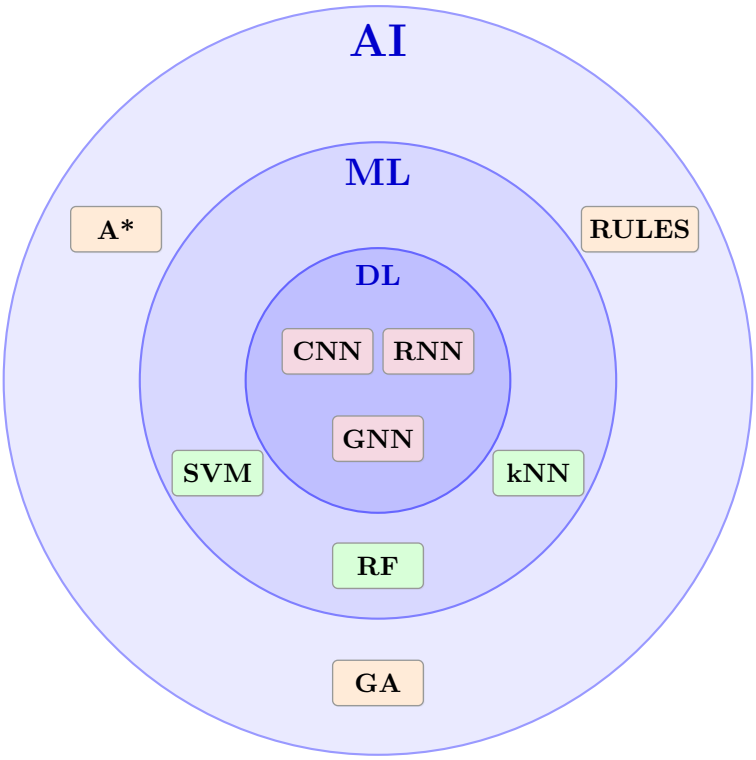


Figure 1.1: Visual overview of AI, ML, and DL methods.

k folds, using one for validation and repeating the process k times to maximize data utility [33].

This balance reflects the variance-bias trade-off [34]. High variance arises if small changes in the input result in significant changes to the model's output. Bias is introduced when complex problems are approximated by simple functions. Consequently, simple models like MLR may be underfitted due to high bias, whereas complex architectures, such as deep NN, are prone to overfitting [34]. Furthermore, utilizing larger datasets mitigates the risk of overfitting by encouraging the learning of general patterns rather than memorizing the full training set. Current best practices often involve pre-training models on large datasets before fine-tuning them for tasks of interest [35].

AI has been successfully applied to solve diverse problems in computer science, including image analysis (often with CNN or more recently ViT) [36], text processing (nowadays with Transformers NN) [37], and social graphs (GNN) [38]. Its application in science is emerging as a vital research area [39], particularly within chemical research. Notably, advancements in areas like protein structure prediction were recognized with the Nobel Prize in Chemistry in 2024, awarded to Hassabis and Jumper [40].

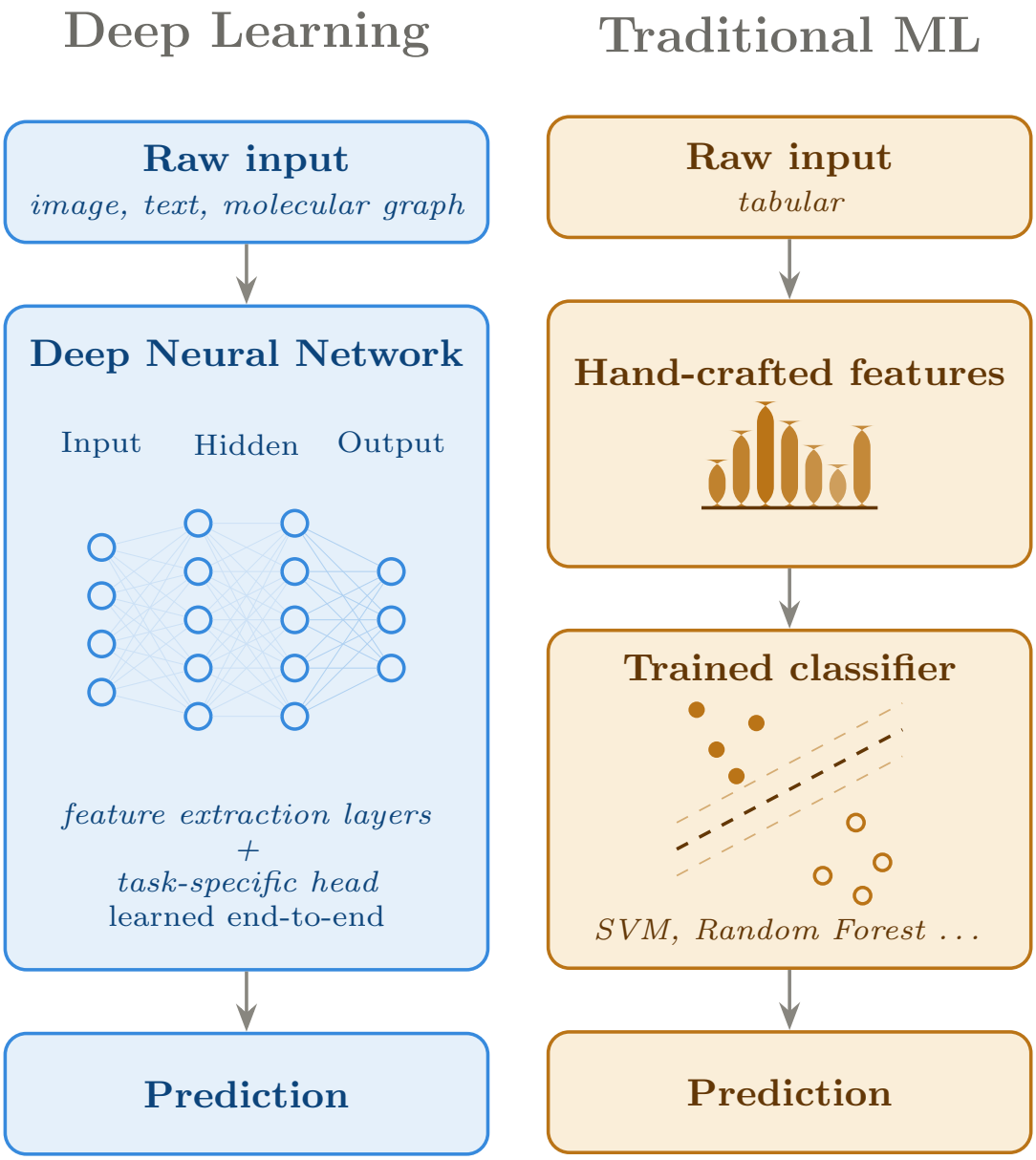


Figure 1.2: Comparison between traditional ML and DL method

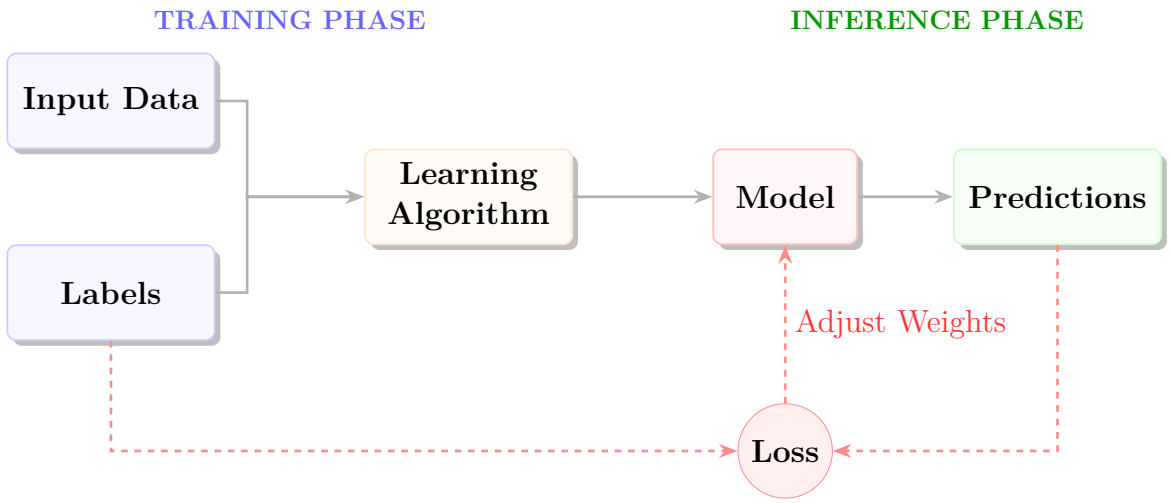
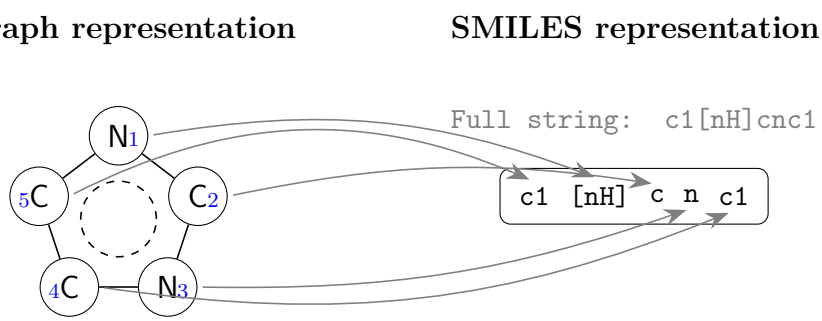


Figure 1.3: Overview of training and inference of AI models

1.2 Chemoinformatics

The integration of AI within the domain of chemistry is primarily investigated through chemoinformatics, a field dedicated to the application of Computer Science (CS) algorithms and methodologies to address chemical challenges. The objectives of chemoinformatics are multifaceted. The primary objectives within chemoinformatics encompass the management of chemical databases, the prediction of molecular properties, and the analysis of chemical transformation outcomes [41].

Chemical data presents a complex medium, necessitating various methods for its storage and representation [42]. One fundamental approach involves utilizing a graph data structure, where atoms are modeled as nodes and chemical bonds are represented by connecting edges. Connectivity information within these graphs is typically encoded using an adjacency matrix. Alternatively, chemical structures can be represented textually through Simplified Molecular Input Line Entry Specification (SMILES) notation. SMILES employs ASCII characters to generate strings that are both unambiguous and human-readable. To mitigate structural ambiguity, SMILES are frequently stored in databases in their canonical form. In this notation, atoms are denoted by their standard periodic table symbols, with lowercase letters indicating the presence of an aromatic ring. Bond types are distinguished by specific symbols: a dash for single bonds, an equals sign for double bonds, and a hash symbol for triple bonds [43]. Furthermore, numbers within SMILES are used to denote the opening and closing points of rings in cyclic compounds. While SMILES is a prevalent choice in molecular property prediction (MPP), SELF-referencIng Embedded Strings (SELFIES) notation offers a more robust alternative for molecule generation, as SMILES strings may not always correspond to a chemically valid molecule [44]. These structural data representations—molecular graphs, SMILES, and SELFIES—are sometimes supplemented by 2D visual images, which are particularly effective for validating data correctness. Conceptually, all these methods can be related back to the molecular graph, which serves as the underlying data structure for connectivity. The conceptual relationship between SMILES and molecular graph representations is illustrated in Figure 1.4.



Schematic by example of imidazole molecule: left = explicit atom graph; right = exemplary linear textual SMILES (tokens linked to atoms).

Figure 1.4: Concepts of SMILES and molecular graphs representations.

Among primary objectives of chemoinformatics, MPP represents a pivotal application, particularly for forecasting the behavior of novel compounds prior to their laboratory synthesis. The successful deployment of MPP frequently necessitates the application of

sophisticated ML algorithms trained on extensive datasets. This methodology typically employs supervised learning to identify pertinent structure-property correlations. The predictive models are often designed to target a diverse range of physicochemical properties (e.g., density, viscosity, surface tension), biological activities, sorption capacities, and activity coefficients [45].

The operational workflow for MPP modeling adheres to a structured procedure [46]. It begins with a comprehensive problem definition, which involves delineating the target variable and the intended use cases of the model, while also determining whether the problem is best suited for regression or classification. Understanding the operational environment is crucial for ensuring appropriate model evaluation. Data acquisition is the subsequent critical step for training the ML model. Given that datasets are often compiled from disparate sources, rigorous data standardization is required. Specifically, molecules represented by SMILES must be normalized to ensure consistent expression of functional groups. Following standardization, the data is partitioned into training and testing subsets, often utilizing random or scaffold-based splitting. A portion of the training data is then reserved as a validation set to optimize model hyperparameters, typically through a CV protocol. The subsequent phase involves either the featurization of data for ML methods or the direct integration of raw graphs or SMILES into DL architectures. Finally, the model undergoes rigorous testing to assess its performance. It is imperative that a realistic test set is employed to simulate the model's efficacy within a practical working environment. The comprehensive overview of the MPP workflow is illustrated in Figure 1.5.

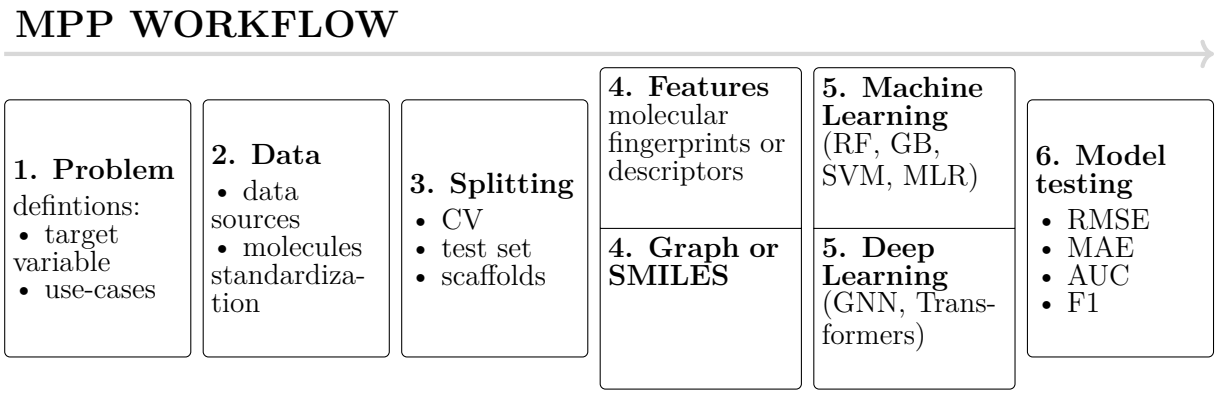


Figure 1.5: MPP workflow overview

The primary objective of property prediction modeling is typically to estimate physicochemical values that would otherwise necessitate empirical measurement within a conventional laboratory setting [47]. This modeling objective is practically achieved by employing a training subset of the dataset to determine the model's parameters, followed by the utilization of a testing subset to rigorously evaluate its predictive performance. This methodology effectively simulates the application of the model to novel experimental data [48]. Given that ML is fundamentally a data-driven paradigm rather than a physics-based approach, the quality and quantity of the data are of paramount importance [49]. Although ML models may be pre-trained or augmented with thermodynamic information, their efficacy in both construction and validation remains critically dependent on the underlying data.

These predictive tasks can be conceptualized as modeling within a high-dimensional vec-

tor space defined by molecular features. Due to the vast combinatorial possibilities of atoms and moieties, this chemical space is inherently multivariate [50]. Several distinct methodologies exist for representing a molecule within this chemical space. One approach involves molecular descriptors (MD), which constructs a numeric vector derived from molecular graphs and/or atomic coordinates. This vector can encode diverse features, including atomic and bond counts, the presence of rings or branches, charge distribution, polarity, molecular weight, and hydrophobicity [51]. A sparser representation is provided by molecular fingerprints (MF), which utilize binary encoding to denote the presence or absence of specific moieties [52]. Alternative methods employ occurrence counters instead of boolean values [53]. Finally, chemical space can be characterized using embedding vectors generated by pre-trained DL models [54].

MFs are frequently employed in the context of assessing the similarity between pairs of chemical compounds. The standard methodology involves the application of the Tanimoto similarity coefficient [55], as conceptually illustrated in Figure 1.6. This procedure commences with the acquisition of MF representations for the two compounds under consideration, followed by the computation of the Intersection over Union (IoU) value. The IoU is mathematically derived by dividing the count of shared bits by the sum of the total bits in both MFs minus the count of shared bits. This similarity metric yields a value ranging from 0, indicating no similarity, to 1, signifying identical compounds [56].

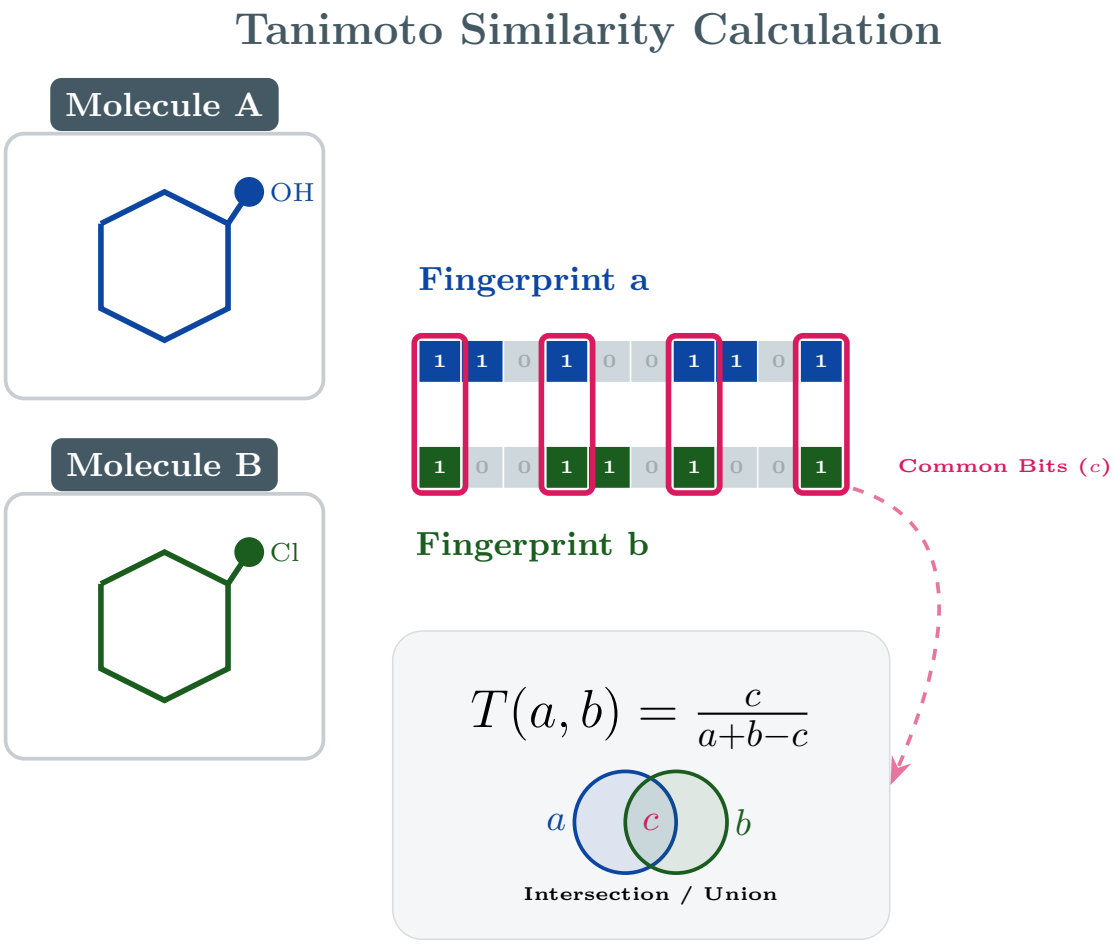


Figure 1.6: Concept of Tanimoto similarity of molecules

1.3 Machine learning in molecular settings

ML demonstrates significant utility in molecular property prediction by enabling the quantitative assessment of the relationship between molecular structures and their associated properties. The application of these predictive models is critically important, as it facilitates substantial savings in time and resources that would otherwise be required for empirical experimentation within traditional chemical laboratories [57]. Furthermore, the integration of virtual screening techniques allows for the prioritization and selection of compounds anticipated to be most promising for specific applications, based on simulated outcomes [58]. Moreover, these models can complement molecular generation strategies, thereby automating and accelerating the overall drug discovery process [59].

The application of ML modeling is particularly warranted when investigating chemical systems characterized by a vast array of possible moiety combinations. Consequently, a highly relevant area of ML research within chemoinformatics involves bimolecular systems, such as ionic liquids (ILs) [60]. ILs are defined as chemical substances comprising two molecular species—a cation bearing a positive charge and an anion bearing a negative charge—that exist in a liquid phase below 100 degrees Celsius [61]. These substances exhibit distinctive physicochemical properties, including low volatility, high melting tem-

peratures, high conductivity, and significant design flexibility [62]. This latter characteristic is of paramount interest to chemoinformatics, as it allows for the precise tailoring of ILs' properties to meet specific functional requirements, leading to their designation as "designer solvents." [63] However, the structure-property relationships governing ILs and their mixtures with solutes are inherently complex and present substantial modeling challenges [64]. These properties are influenced by a multitude of factors, including the specific nature of the cation, anion, and solvent composition, which precludes the development of a universally applicable predictive model [65]. Computational methodologies, including ML, can address several modeling objectives related to ILs, such as predicting melting points, solubility, conductivity, viscosity, and toxicity [66]. The utility of ILs is extensively documented across various disciplines, including analytical chemistry (e.g., SPME extraction), separation processes (e.g., carbon dioxide sequestration), physical chemistry (solubility studies), organic chemistry (catalysis), and chemical engineering. Furthermore, the application of ILs in sorption systems necessitates the development of models capable of describing multimolecular systems, which consist of the bimolecular solvent and the solute molecule [67].

The application of ML methodologies for predicting the properties of ILs is inherently complex, primarily due to the significant heterogeneity present in both the underlying databases and the chemical space of the ILs themselves [68]. Given that ILs are composed of distinct electrically charged species (cations and anions), their chemical diversity is substantial, a complexity further exacerbated when experimental parameters, such as temperature and pressure, must be integrated for accurate predictive modeling—a necessity particularly relevant in sorption modeling [69]. A critical impediment to robust ML development is the scarcity of high-quality training data, a common challenge stemming from the unique chemical nature of ILs compared to conventional drug-like organic compounds [70]. Furthermore, the applicability of pre-trained models is severely limited by the disparate chemical characteristics between ILs and typical organic molecules [71]. Data quality presents another major hurdle; discrepancies arising from varied measurement techniques across different laboratories, coupled with limited methodological reproducibility, necessitate a rigorous standardization process to mitigate noise that could compromise ML model performance. Additional challenges include the potential influence of trace contaminants—such as water, other ions, acetone, or ethanol—introduced during synthesis, which may alter ILs' properties, yet these factors are frequently omitted from published literature [72]. Despite these challenges, the vast combinatorial potential of cation-anion pairings renders ILs highly suitable for virtual screening aimed at identifying optimal solvents, positioning them as a compelling subject for chemoinformatics investigation [73]. Consequently, this thesis adopts ILs as the foundational subject of study.

Chapter 2

Related works

2.1 Classical molecular property prediction

Classical ML methodologies are predicated on the concept of chemical space, which facilitates the exploration of diverse moieties and scaffolds through multidimensional vector spaces. This chemical space is typically characterized by MD, MF, or similar predefined numerical representations [74]. A wide array of molecular representations has been proposed within this context, ranging from basic functional group counts to more sophisticated methods. For instance, popular MF approaches, such as the Morgan algorithm, utilize hash functions to generate fixed-length bit vectors that encode the presence of specific substructures [75]. Similarly, ISIDA descriptors quantify the frequency of subgraph occurrences within a molecular graph [76], [77]. Many studies rely on descriptors derived from molecular graphs (e.g., 2D RDKIT descriptors), which include features such as atom or rotatable bond counts, ring counts, the sum of shortest path distances, and connectivity indices [78]. Furthermore, the incorporation of three-dimensional atomic coordinates (as implemented in software like Dragon of AlvaDesc) allows for the expansion of the MD feature set [79], [80]. Common 3D descriptors include molecular volume, surface area, shape indices, and 3D autocorrelation functions [42], [81]. Advanced evaluations of atomic coordinates, often coupled with quantum mechanical calculations, enable the utilization of partial atomic charges or polarizability as MD [82]. More recently, the field has shifted toward vectorizing molecular data using embeddings learned through GNN or transformer-based NN architectures [54], [83].

ILs present a particularly compelling subject for chemical-space analysis due to their bimolecular nature, which generates a vast and structurally heterogeneous design space [84]. When conceptualized as bimolecular systems, the chemical space of ILs is defined by the structural descriptors of both the anionic and cationic components. In practical applications, this vector space is characterized by high dimensionality, often requiring hundreds or thousands of descriptors per ion [85]. The primary objective of model users is typically to explore specific regions of this chemical space that correlate strongly with high values of the target property. However, when the ILs are involved in a multi-component system, such as in a sorption process, the modeling challenge shifts from characterizing the ILs in isolation to characterizing the entire system. Consequently, in sorption modeling the chemical space of the solute must also be considered [86]. These multifaceted challenges are conceptually illustrated in Figure 2.1.

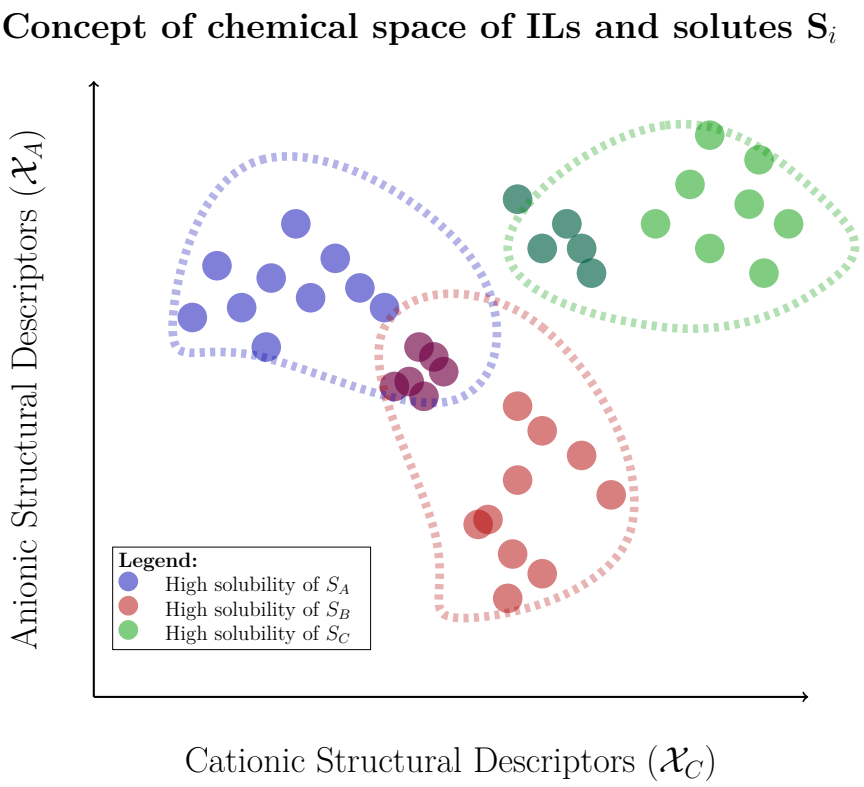


Figure 2.1: Concept of chemical space in multimolecular systems

In the domain of molecular modeling, various datasets are frequently employed as standard benchmarks for validating novel ML methodologies. The MoleculeNet benchmark [87] is a comprehensive collection of datasets spanning applications in quantum chemistry, physical chemistry, biophysics, and physiology. Specifically, the QM9 benchmark [87] provides data on molecular properties derived from quantum mechanical calculations, including target variables such as solubility and lipophilicity within the physical chemistry domain, as well as compound activity against HIV or toxicity. These datasets exhibit a wide range of scales, encompassing between 600 and 437,000 compounds. Furthermore, the FS-Mol datasets [88] offer diverse data pertaining to activity against protein targets, while the TDC ADMET [89] database comprises 22 endpoints relevant to the prediction of absorption, distribution, metabolism, excretion, and toxicity. However, for more specialized applications, such as the modeling of multimolecular systems, the establishment of universally accepted standard benchmarks remains an unresolved challenge.

This situation is more perplexed with ILs, however. Within the scope of classical modeling paradigms, datasets are conventionally organized in a tabular structure, wherein each row represents a unique molecule. However, this conventional organization becomes significantly more complex when addressing ILs. Certain studies exhibit a one-to-many relationship, wherein a single compound is associated with multiple records within the dataset. This structural characteristic frequently arises when the target property is influenced by external variables such as temperature, pressure, or other environmental conditions. Specifically concerning ILs, this pattern is common in datasets pertaining to sorption or physicochemical properties [90]. This structural complexity is illustrated in Figure 2.2.

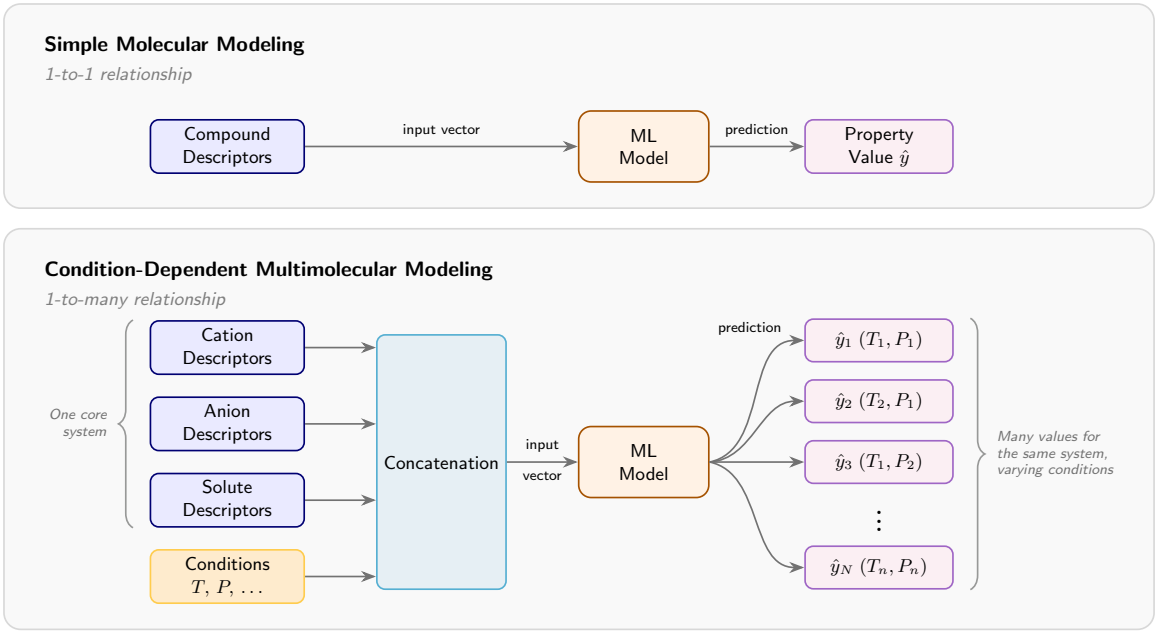


Figure 2.2: Comparison between simple molecular modeling and multimolecular modeling.

Various datasets detailing the physicochemical properties of ILs are documented within the relevant literature. Specifically, a series of publications by Paduszyński introduced three distinct databases covering the density (approximately 40,000 data points), viscosity (around 20,000 data points), and surface tension (approximately 6,000 data points) of ILs. Due to their recent publication and comprehensive scope, these databases serve as established benchmarking standards for physicochemical properties modeling within the ILs domain. Furthermore, extensive toxicity data for ILs are available in large-scale databases; for instance, the ILTox database encompasses 1,180 ILs and includes various endpoints such as E. coli, AChE, IPC-81, and Vibrio fischeri. Cytotoxicity data for over 1,200 ILs are also widely available in the literature. Regarding sorption characteristics, the ILThermo [91] database provides a substantial collection of data, including approximately 40,000 data points pertaining to infinite dilution activity coefficients (IDAC).

Molecular descriptors, typically organized into tabular datasets, serve as the input features for various ML algorithms, including MLR, Multilayer Perceptron (MLP), and GB. Among these, decision tree-based algorithms have demonstrated significant prevalence in this domain. These methods possess the theoretical capacity to be effectively fitted even when data availability is constrained, leading to extensive research into their applicability for relatively small datasets. For instance, Svetnik et al. [92] demonstrated the superior performance of the GB algorithm compared to other structure-activity modeling techniques across both small and large datasets. Similarly, Shi et al. [93] utilized this approach to predict acute exposure guideline levels based on a collection of 90 aliphatic compounds, while Wiriyarattanakul et al. [94] applied it to a dataset comprising 45 anti-inflammatory compounds. Alternatively, linear models have also been extensively investigated due to their high degree of interpretability and minimal requirement for large training sets. Ambure et al. [95] developed specialized software for the development of linear models using small molecular datasets, and Lavado et al. [96] employed this methodology to model toxicity using a dataset of 54 compounds. Furthermore, Zhang et al. [97] successfully

applied linear models even with a limited sample size of 48 compounds.

The predictive modeling of physicochemical properties in ILs has been a subject of significant investigation, as evidenced by the trends observed in the literature [66]. Consequently, this section provides a concise overview of selected studies in this domain. Paduszyński employed a group-contribution methodology to model the properties of ILs, wherein physicochemical features are derived from predefined atomic groups and their respective occurrences. This approach demonstrated high efficacy in predicting density, achieving low error metrics (MARE ≈ 0.015), and exhibited strong predictive capability for surface tension, with an R^2 approaching 0.99 and $MARE \approx 0.085$. However, the modeling of viscosity proved more challenging, yielding comparatively lower accuracy (best $R^2 \approx 0.84$; $MARE \approx 0.377$). Similar challenges were noted by Billard et al. [98], who reported an R^2 of approximately 0.7 on their test set. Furthermore, Varnek et al. [99] utilized molecular representations, including ISIDA descriptors, alongside SVM and NN to predict the melting points of ILs, achieving robust predictive power with an error of approximately 40 degree Celsius on novel compounds. Finally, Venkatraman et al. [100] demonstrated the superiority of ML techniques, such as RF and GB, over traditional COSMO-RS methods for the prediction of ILs melting points.

The application of ML models to predict sorption phenomena within ILs has become a significant area of research. Specifically, the solubility of gases, particularly carbon dioxide, has been extensively investigated, with Jian et al. [69] employing GNN architectures for this purpose. Beyond gaseous species, the interactions of ILs with other liquid phases have also been explored; for instance, Klimenko et al. [90] utilized decision tree-based ML methods to model solubility in water-ILs systems. Furthermore, the prediction of solid solubility has been addressed, as demonstrated by Euldji et al. [101], who successfully modeled the solubility of pharmaceuticals within ILs.

Prior research has also extensively addressed the challenge of modeling ILs toxicity through various computational approaches. For instance, Kang et al. employed multiple linear regression, incorporating sigma profiles, to predict EC_{50} values in leukemia rat cells [102]. Expanding the scope, Abdellatif et al. utilized a SVM model to assess toxicity within the IPC-81 dataset, drawing upon a comprehensive collection of 304 ILs [103]. Similarly, Semenyuta et al. developed predictive models using toxicity data from Daphnia magna (75 ILs) and Danio rerio (99 ILs), demonstrating robust performance with R^2 values approaching 0.80 [104]. The computational modeling of toxicity is often constrained by the time- and cost-intensive nature of experimental protocols, which inherently limits the available dataset size for ML applications. To circumvent these limitations, several studies have leveraged advanced ML techniques. Yan et al. [105] utilized the GB algorithm for modeling ILs toxicity, while Shan et al. constructed ML models by integrating data on E. coli, Acetylcholinesterase (AChE), and IPC-81 toxicity. Their implementation of a RF algorithm, operating on two-dimensional MD data, yielded highly effective performance [106]. Furthermore, Fan et al. applied tree-based algorithms to model the inhibitory effects on Vibrio fischeri [107], and Wang et al. focused on neural networks and SVM to formulate predictive models for IPC-81 toxicity [108]. In a separate analysis, Feng et al. examined the ILTox database and observed that utilizing a minimum of 140 individual compounds could potentially lead to a plateau in the loss function curve when plotted against the number of training instances. Nevertheless, their investigation did not explore specialized solutions tailored for few-shot molecular property prediction [109].

2.2 Neural networks in molecular settings

Various types of NN have been employed in the field of molecular modeling. Specifically, GNN are frequently utilized in MPP studies. The graph data structure is inherently suited for chemical modeling, as atoms can be represented as nodes and chemical bonds as edges. By leveraging graph priors, these models can directly incorporate information regarding chemical connectivity. For instance, the chemprop library [110] facilitates message passing between nodes (Atom Message Passing) or edges (Bond Message Passing) within a graph framework. Building upon this library, Burns et al. [111] recently proposed the ChemMeleon foundation model. In contrast, ChemBERTa employs a different architecture, utilizing a pre-trained transformer NN. This approach conceptualizes chemistry as a formal language, enabling the application of sequence modeling NN to encode SMILES strings. Both ChemMeleon and ChemBERTa were pre-trained on extensive datasets of compounds (up to approximately one million) for the task of predicting classical MD values. Consequently, the mean squared error between the predicted and actual descriptor values was adopted as the loss function during the pre-training phase. The objective of the pre-training phase is to enable the model to acquire robust feature extraction capabilities. This stage can be designed around various tasks, such as predicting classical MD values (as it was explained) or reconstructing masked fragments within the SMILES string. Following this, the resulting base model can be fine-tuned for a specific downstream application. A key advantage of this two-step protocol is that fine-tuning typically requires significantly less data than training a DL model from scratch. This methodology is visually represented in Figure 2.3.

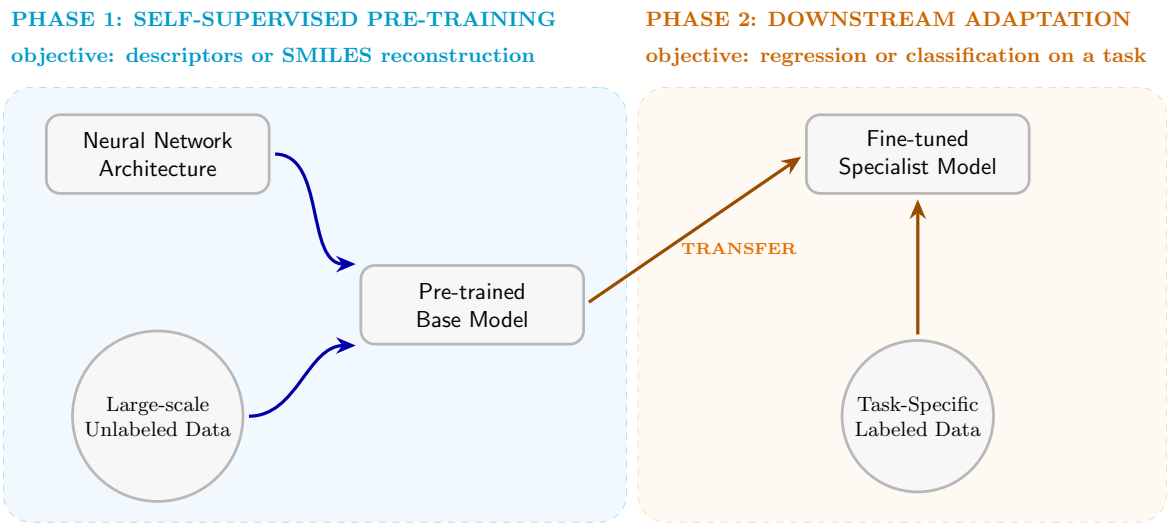


Figure 2.3: Pre-training and fine-tuning of deep neural networks

Pre-trained models can be also derived through the application of autoencoder (AE), a class of NN architectures designed to compress input data and subsequently reconstruct the original information from a latent representation. While the conventional AE employs a deterministic encoder-decoder framework, the variational autoencoder (VAE) introduces a probabilistic paradigm suitable for generative modeling. Unlike standard AE, which maps inputs to a single fixed vector, the VAE encoder outputs a probability distribution over latent variables. Furthermore, the conditional variational autoencoder

(CVAE) extends this capability by incorporating auxiliary information, such as class labels, to condition the generative process. These variants of AE have been successfully applied to chemical informatics. For instance, Gómez-Bombarelli et al. [112] developed a VAE specifically for SMILES reconstruction, demonstrating that the resulting latent vector serves as a comparable featureizer to traditional methods like MD or MF. This approach facilitated the generation of novel molecules optimized for specific properties. However, a significant limitation of this method is the potential for generated SMILES strings to represent chemically invalid structures or violate valency laws. To address this, Jin et al. [113] proposed utilizing a tree structure during multi-step optimization, thereby enforcing the generation of chemically valid graphs at each step. Similarly, Kusner et al. [114] employed a tree-based parsing of SMILES to extract grammatical rules. By operating on these structural rules rather than raw characters, their method ensured the generation of syntactically valid molecules.

The application of deep architectures to ILs has gained significant traction in recent years. For instance, Sadaghiyanfam et al. employed a hybrid methodology integrating ChemBERTa embeddings, MD, and MF for toxicity prediction [115]. Their analysis demonstrated that deep neural network embeddings exerted a greater influence on predictive accuracy compared to traditional molecular representations. More recently, the potential of GNNs has been extensively explored in the context of ILs property prediction. Specifically, GNNs have been utilized to model carbon dioxide solubility [69] by structuring cation and anion representations within a unified matrix, where cation graphs occupy the upper left and anion graphs the lower right, with null values filling the remainder. Concurrently, a distinct methodological approach was adopted by another study [116] to investigate the modeling of activity coefficients for solutes in ILs. In contrast to the matrix-based strategy, this approach involved concatenating the cation and anion vectors only subsequent to the graph convolutions, immediately preceding the final fully connected layers.

Beyond the application of GNN and transformers, several other architectures have been employed in the prediction of ILs properties. For instance, Fan et al. [117] utilized a CNN approach on a dataset comprising 155 ILs to predict toxicity, achieving highly satisfactory metrics ($R^2 = 0.965$). However, the applicability of DL in this study was constrained by the limited dataset size and the small test set (40 records). In a related work, Makarov et al. [118] investigated the prediction of ILs melting points, employing a variety of algorithms, including CNN and transformers-CNN.

Further research has explored generative models. Ren et al. [119] integrated gated recurrent unit (GRU) layers within an NN framework for a CVAE, conditioning the AE on a label indicating the likelihood of ion structure formation in ILs. This methodology subsequently led to a state-of-the-art model for predicting ILs melting points ($R^2 = 0.75$). Similarly, Chen et al. [73] developed a GRU-based VAE for the reconstruction of ion SMILES, which they fine-tuned for predicting carbon dioxide solubility in ILs. The utility of CNN architectures in AE has also been demonstrated for SMILES reconstruction, as seen in the work of Liu et al. [120], where 1-D convolution operations were utilized. Their AE-based solution proved superior to classical descriptors when predicting carbon dioxide solubility. Furthermore, the efficacy of CNN-based VAE models has been evaluated using one-hot encoding of SMILES strings [121].

As a result, the application of NN in molecular contexts remains relatively underexplored,

primarily due to the demonstrated superiority of GB in many instances. To address this deficiency, the Independent Feature Mapping (IFM) method was introduced [122]. This approach enhances the molecular representation capabilities of neural networks by independently transforming each feature using sine and cosine wave functions. This mechanism enables the system to capture fluctuations across a wide spectrum of frequencies, thereby facilitating a more robust understanding of intricate dependencies, such as the relationship between subtle chemical modifications and significant variations in molecular properties.

2.3 Handling data limitation

Despite the aforementioned challenges, the inherent data limitations persist as a significant constraint for NN and other structure-property models. This deficiency may, however, be partially mitigated through the application of pre-training utilizing synthetic data. Specifically, Vermeire et al. [123] demonstrated that pre-training GNNs on COSMO-RS derived data can be highly beneficial when employing DL methods to model small experimental datasets.

A more targeted methodology could leverage meta-learning algorithms, which provide a robust framework for addressing data scarcity by enabling rapid model adaptation. Among these, the Model-Agnostic Meta-Learning (MAML) algorithm is widely employed due to its versatility across various NN architectures. As introduced by Finn et al. [124], MAML is designed to address diverse problems, including regression, classification, and reinforcement learning, fundamentally operating on the principle of "learning how to learn." This approach employs a two-loop protocol: an inner loop facilitates adaptation to a novel task using a limited adaptation set, while the outer loop assesses the generalization accuracy of this adaptation across a batch of tasks. However, the inherent complexity of this protocol often imposes significant constraints on the scalability of MAML. Consequently, there has been a growing necessity for more computationally efficient, task-agnostic learning paradigms. The Reptile method addresses this by iteratively sampling tasks, executing stochastic gradient descent (SGD) for each task, and subsequently updating the NN weights. By omitting the reliance on second-order derivatives, the Reptile algorithm achieves faster computational performance [125].

The application of the MAML method has been extensively explored within the domain of chemoinformatics. Pappu et al. demonstrated that while GNN typically struggles against fingerprint-based methods, the integration of meta-learning significantly boosts GNN performance, surpassing both pretraining and fingerprint-based approaches in low-data scenarios [126]. This aligns with the findings of Nguyen et al., who established that meta-initializations yield superior outcomes compared to alternative deep learning methodologies when adapting to restricted datasets [127]. Dou et al. identified MAML as a crucial algorithm for enabling low-data machine learning in chemical applications [128]. Building upon this, Guo et al. integrated MAML with a GNN architecture to develop predictive models for chemical toxicity assessment. Their investigation utilized the Tox21 dataset and evaluated performance under both 1-shot and 5-shot fine-tuning conditions, where the latter refers to adapting the classification model using five compounds per task. The results demonstrated that leveraging MAML significantly improved model performance, raising the AUC score from 75.83 to 76.87 in the single-shot scenario and from 77.18 to 78.02 in the five-shot setting [129]. Similarly, Ju et al. investigated the Tox21 and

ToxCast datasets using the MAML-GNN framework. Their findings revealed that even with a minimal adaptation set of only ten samples, they achieved impressive AUC metrics of 80.21 for Tox21 and 66.79 for ToxCast [130]. In a different context, Schlender et al. employed a dataset comprising 24,816 aquatic toxicity values (LC50) across 351 species, integrating MAML into both single-task and multi-task modeling paradigms [131]. Lv et al. confirmed MAML's superiority in regression tasks involving quantum mechanical parameters and further established its leading performance among various few-shot algorithms for predicting drug-drug interactions [132], [133]. Similarly, Qian et al. highlighted the effectiveness of an attention-based GNN pre-trained via MAML on CHEMBL datasets, noting its exceptional performance in classification tasks, particularly on the MoleculeNet benchmark [134].

However, the application of meta-learning is not without challenges. Meng et al. identified few-shot MPP as a critical yet difficult area for improving virtual screening success rates, while also cautioning that meta-initialization risks, such as memorization of query samples or overfitting to training tasks, can impede effective test task adaptation [135]. Furthermore, Kotter et al. scrutinized how the inclusion of similar data during meta-training influences the fine-tuning of models grounded in biological activity datasets [136]. Notably, the techniques employed in these studies relied heavily on neural network models, a choice that warrants critical consideration as neural networks are not always the preferred methodology in structure-activity modeling [122].

Chapter 3

Scope and goals of the thesis

Drawing upon the preceding synthesis of extant literature, it is evident that ML possesses substantial potential for advancing scientific inquiry within chemical domains. Despite the prevailing consensus that ML necessitates extensive datasets for effective training, this assumption has spurred significant investigation into few-shot modeling techniques. Furthermore, the utility of ML is not exclusively constrained by data availability; however, a comprehensive understanding of its precise applicability remains underdeveloped. This serves as the foundational premise for the present investigation.

The scope of this thesis focuses on low-data learning within the context of multimolecular systems. This specific constraint facilitates a comprehensive investigation into the application of ML in chemical science, particularly in scenarios that mimic data-scarce environments—a critical aspect of scientific discovery. The relevance of this scope is underscored by the fact that few-shot learning remains significantly underexplored in molecular modeling, especially when dealing with heterogeneous data such as that pertaining to ILs. Although some research on few-shot molecular system modeling exists, the literature on this subject remains notably sparse. Existing studies have predominantly applied few-shot learning to simpler molecular systems, often within the framework of common MPP benchmarks, which may not be optimally suited for settings involving numerous related tasks. Furthermore, the prevailing literature tends to treat the MAML algorithm as a standard, unmodified tool, lacking critical discussion regarding potential domain adaptations. Crucially, the inherent limitations of MAML have neither been systematically identified nor adequately addressed. Consequently, this research area necessitates further rigorous examination.

The scope of this thesis is structured to achieve the following objectives:

G1. To critically evaluate the inherent limitations of both classical and deep learning methodologies when applied to the prediction of ILs' properties.

G2. To ascertain the specific added value and operational constraints associated with the implementation of few-shot learning techniques.

G3. To propose modifications to existing algorithms or training protocols aimed at significantly enhancing model accuracy for particular problem domains.

The central hypothesis of this research posits that both conventional and deep ML methodologies demonstrate significant utility in the modeling of multimolecular systems

characterized by limited data availability, provided they are appropriately engineered, particularly when leveraging meta-learning techniques.

A comprehensive understanding of the ML methodologies applicable to low-data regimes is a fundamental prerequisite for advancing this field. While initial insights derived from simpler molecular systems were anticipated to generalize to multimolecular systems, certain aspects appear to have been insufficiently addressed. Specifically, it is imperative to investigate the influence of chemical space commonalities between tasks, particularly in context of few-shot learning protocols integrated with meta-learning frameworks. Furthermore, the research should move beyond merely analyzing the data encapsulated within tasks to examine how the inherent similarity between tasks themselves affects model performance. Additionally, the selection of the adaptation set and its subsequent impact on adaptation uncertainty warrant rigorous exploration. To properly evaluate these issues, the utilization of baseline controls, encompassing both classical and deep learning tools, is essential. Comparative evaluation against these conventional modeling strategies not only highlights the inherent strengths of the examined few-shot learning methods but also facilitates a deeper analysis of their advantages and limitations. This critical examination provides a foundation for developing potential improvement strategies, which may involve either refining the training strategies of established algorithms or modifying original training procedures to achieve desired performance enhancements.

Consequently, several research questions were identified as relevant for further study:

Q1. What is the efficacy and relevance of integrating classical and deep learning methodologies in addressing contemporary chemical research challenges, such as carbon dioxide absorption and lignin isolation?

Q2. To what extent does the incorporation of rule mining and physics-based descriptors enhance the predictive capability of ML models for bimolecular systems?

Q3. What optimal implementation strategies are required for GNN models to achieve peak performance in predicting the properties of ILs?

Q4. How do the quantity and quality of available data influence the predictive accuracy and robustness of GNN models when predicting ILs' properties?

Q5. How does few-shot meta-learning compare to traditional ML paradigms in the context of modeling complex bimolecular systems?

Q6. Is the utilization of low-quality simulated data for pre-training meta-learning models a viable strategy for addressing data scarcity in low-data learning scenarios?

Q7. What are the inherent limitations of meta-learning approaches concerning the similarity between tasks?

Q8. How does the similarity of chemical spaces between tasks constrain the applicability and performance of meta-learning models?

Q9. Can modifications to the adaptation (inner loop) phase of meta-learning serve as an effective strategy to improve performance and reduce the standard deviation of metrics, particularly regarding adaptation set selection?

Q10. Is the integration of a task similarity factor, approximated via Tanimoto coefficients, a beneficial strategy for enhancing the modeling of more challenging tasks within

meta-learning algorithms?

The research objectives and the resolution of the posed questions were achieved through a multi-stage methodological approach. The initial phase involved preliminary investigations into the application of ML for multimolecular systems, focusing on two distinct domains: lignin isolation and carbon dioxide capture. These two datasets were selected based on their significant relevance to contemporary chemical research. Subsequently, the second phase addressed the implementation of deep graph neural network training protocols to resolve specific challenges encountered during the modeling of multimolecular system properties. The third stage utilized few-shot learning modeling, leveraging prior expertise derived from both classical (descriptors-based) and deep learning methodologies. The alignment between the research steps, defined goals, and specific research questions is visually delineated in Figure 3.1.

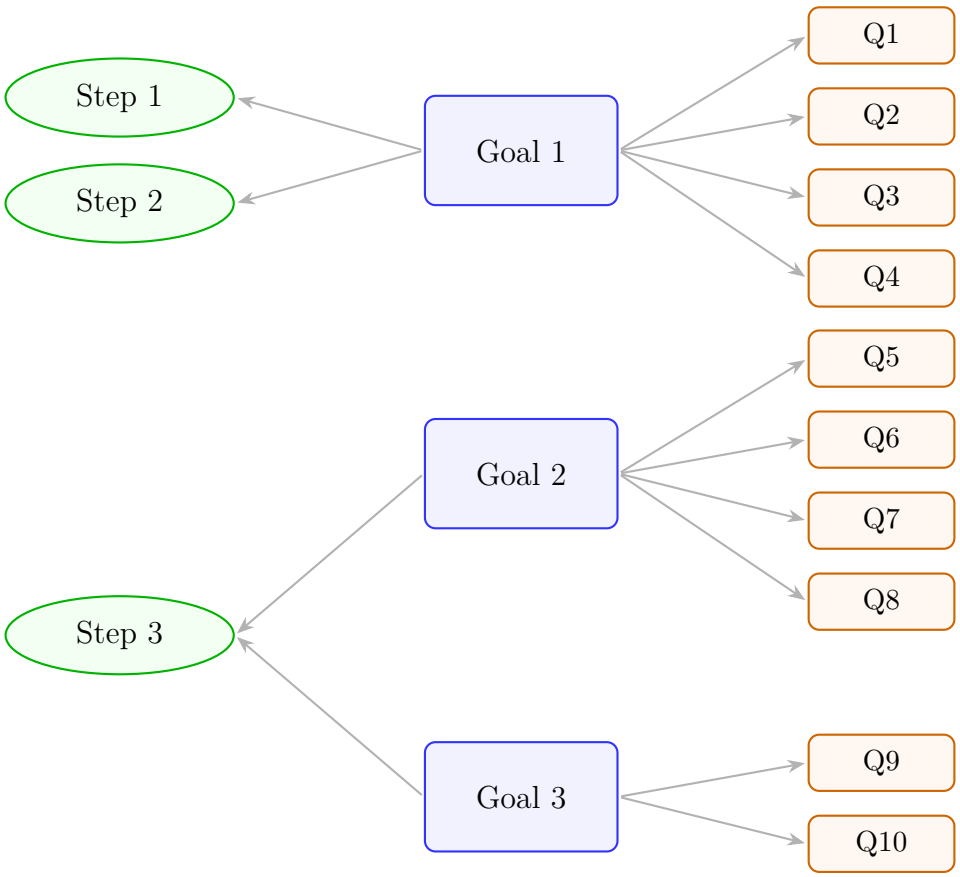


Figure 3.1: Relationship between research steps, goals, and questions

The central aim of this research is to advance the computational methodologies currently employed within the domain of ML for the modeling of ILs systems. This thesis investigates this intersection from the perspective of chemistry and CS, utilizing various ML techniques to address specific research questions. Unlike studies that focus on narrow ML applications in chemistry, this work specifically targets broader chemoinformatics challenges. Through a comparative analysis of algorithms and models, the suitability of the adopted approach is validated. This investigation holds significant relevance for both CS and chemistry, offering case studies focused on low-data learning. A major challenge addressed is the inherent heterogeneity of chemical data, which stems from variations in molecular systems and experimental conditions. Within the framework of meta-learning,

this research provides insights into how task similarity influences the performance of MAML models. Furthermore, the architectural or algorithmic adaptations proposed offer unique contributions to CS, while the prediction of ILs properties remains highly pertinent to chemical research. Finally, the practical implications of this work are substantial across numerous industrial sectors, including the extraction and separation of mixtures, gas sorption (such as carbon dioxide), and purification processes like crystallization (e.g., in the production of active pharmaceutical ingredients).

This thesis advances the AI research community by addressing a critical domain gap. Although significant strides have been made in AI development - particularly concerning text, image, and sound modalities - the application of these technologies to chemistry remains an understudied area. Consequently, the domain-specific adaptations proposed herein aim to augment the existing body of literature.

Chapter 4

Methods

This chapter provides an overview of the methodological frameworks adopted throughout this research. Given that the experimental design was contingent upon the specificities of the datasets and the nature of the research questions addressed, the protocols were customized for each application, precluding the formulation of a single, standardized procedure. The implemented protocols are described in the Methods sections of the individual publications, and the associated source code is available in the corresponding repositories.

4.1 Datasets collection, preprocessing, and analysis

Addressing the research objectives necessitates the selection of datasets specifically curated to align with the scope of the investigation. These datasets encompass a broad spectrum of properties and applications pertaining to ILs. Specifically, the following datasets were utilized:

- DS1 - (a) Henry's Law Constant of ILs - carbon dioxide systems (data source: Eichenlaub et al. [137]), (b) data on Henry's Law Constant of ILs and other gasses collected from ILThermo [91]
- DS2 - lignin isolation yield with ILs (data source: Baran et al. [138])
- DS3 - physicochemical properties of ILs - density (DS 3a), viscosity (DS 3b), and surface tension (DS 3c) (data source: Paduszynski [68], [139], [140])
- DS4 - gas solubility in ILs expressed with mole fraction (data sources:Jian et al. [69], Dong et al. [91])
- DS5 - toxicity of ILs (data source: [105])
- DS6 - infinite dilution activity coefficients in ILs - solutes systems (data source: [116])

The datasets DS1 and DS2 pertain to topics that have recently garnered significant attention, particularly within the field of environmental science. In contrast, DS3 provides a compilation of fundamental physicochemical properties of ILs. The remaining three datasets (DS4, DS5, and DS6) are specifically designed for meta-learning modeling. DS4 and DS6 represent solvent-solute systems, where a predictive task is defined by the solute. The primary objective in these systems is to forecast the solubility of emerging chemical

entities, thereby addressing the challenge of optimizing solvent selection for specific solutes. DS5, conversely, encompasses information across multiple toxicity endpoints, each representing a distinct task. Consequently, the modeling approach utilizing DS5 aims to investigate whether toxicity against less extensively studied endpoints can be reliably estimated even when data is limited. The rationale underpinning the selection of these datasets is detailed in Table 4.1.

Table 4.1: Summary of the utilized datasets.

Dataset	Dataset Size	Justification for Selection
DS1a	76 records	A small dataset relevant for low-data settings, providing strong baselines [137].
DS1b	84 records	An augmentation of DS1a, incorporating data from various other gases [91].
DS2	108 records	A small dataset intended to demonstrate the ML potential for complex process modeling.
DS3a	ca. 40,000 records	A large dataset appropriate for the evaluation of DL methods.
DS3b	ca. 20,000 records	A large dataset appropriate for the evaluation of DL methods.
DS3c	ca. 6,000 records	A large dataset appropriate for the evaluation of DL methods.
DS4	ca. 15,000 records	A large dataset appropriate for the evaluation of both DL and MAML methods.
DS5	ca. 3,500 records	A relatively small dataset that allows for the isolation of chemical space impact (data lacks temperature dependency).
DS6	ca. 40,000 records	A large dataset with a diverse collection of tasks, suitable for meta-learning in sorption settings (data includes temperature dependency).

The DS1 dataset was further expanded through the simulation of data covering 4,617 distinct combinations, involving 81 unique cations and 57 unique anions. These simulations were conducted within a system comprising six gases: carbon dioxide, carbon monoxide, ethylene, hydrogen, nitrogen, and propylene. All computational procedures were executed utilizing the AMS 2023.104 COSMO-RS package [141].

To ensure comparability across heterogeneous data sources, a comprehensive normalization procedure was implemented. Specifically, molecular structures, represented via SMILES, underwent standardization utilizing the established protocols within the RDKit library [142]. This standardization is critical for maintaining consistent molecular representations, thereby facilitating accurate feature extraction. Prior to subsequent analytical stages, standard preprocessing techniques, including deduplication and filtering, were rigorously applied. For non-decision tree-based algorithms, target features were nor-

malized using min-max scaling to constrain their range between 0 and 1. Outlier detection was initially performed using conventional statistical metrics, specifically the Interquartile Range (IQR), Mean Absolute Deviation (MAD), and Z-score, applied to continuous variables such as density, viscosity, and surface tension. When these statistical methods proved insufficient, the Isolation Forest algorithm was utilized to identify anomalies within high-dimensional datasets, leveraging its capacity to isolate outliers based on their ease of separation within an ensemble of decision trees.

Feature selection within the traditional ML framework was executed by utilizing the feature importance rankings derived from the GB model. The Arithmetic Residuals in K-Groups Analysis (ARKA) dimensionality reduction technique is predicated on the hypothesis that the contribution of individual features varies significantly across distinct ranges of the response (target) variable. This novel methodology was employed contingent upon its demonstrated superiority over a simple feature selection [143].

Data analysis was further augmented with the rule mining methods. The CN2 algorithm refines if-then rules in an iterative manner based on entropy or weighted accuracy measures. The CN2 rules were obtained using the Orange Data Mining package for Python [144]. The rules were evaluated using support (fraction of records following the rule), confidence (proportion of data points that follow the rule and are associated with a correct label by it), and by comparison of the general distribution and the distribution of records following the rule with chi-square test of goodness of fit.

4.2 Validation protocols

The ML protocol necessitates the utilization of data for both model weight acquisition and subsequent validation [34]. Consequently, the dataset is typically partitioned into training, validation, and test subsets. The training set is employed to determine the model's weights, while the validation set serves to optimize hyperparameters—non-trainable parameters that must be specified prior to model training. In this investigation, hyperparameter optimization was achieved through a random search protocol, which iteratively selects hyperparameter configurations to identify the optimal set based on performance metrics derived from the validation set [30].

The selection of the validation strategy can be complex and is contingent upon the dataset size. For smaller datasets, five-fold cross-validation (CV with $k = 5$) was implemented. Conversely, for larger datasets, a random selection of the validation set was preferred over cross-validation due to the significant reduction in computational overhead required for training large-scale models [30]. In these instances, the validation set was allocated approximately 10–15% of the total data.

Standard practice in ML modeling dictates that the final model be evaluated using a dedicated test set. The primary objective of this test set is to simulate the model's performance in a real-world operational environment [30]. The methodology for data partitioning is illustrated in Figure 4.1, and the specific configuration of the test set is detailed in Table 4.2.

The application of meta-learning necessitates specific considerations regarding evaluation protocols. Given that the fundamental objective of meta-learning is to optimize the learning process itself, the model must be evaluated on a set of held-out tasks. These

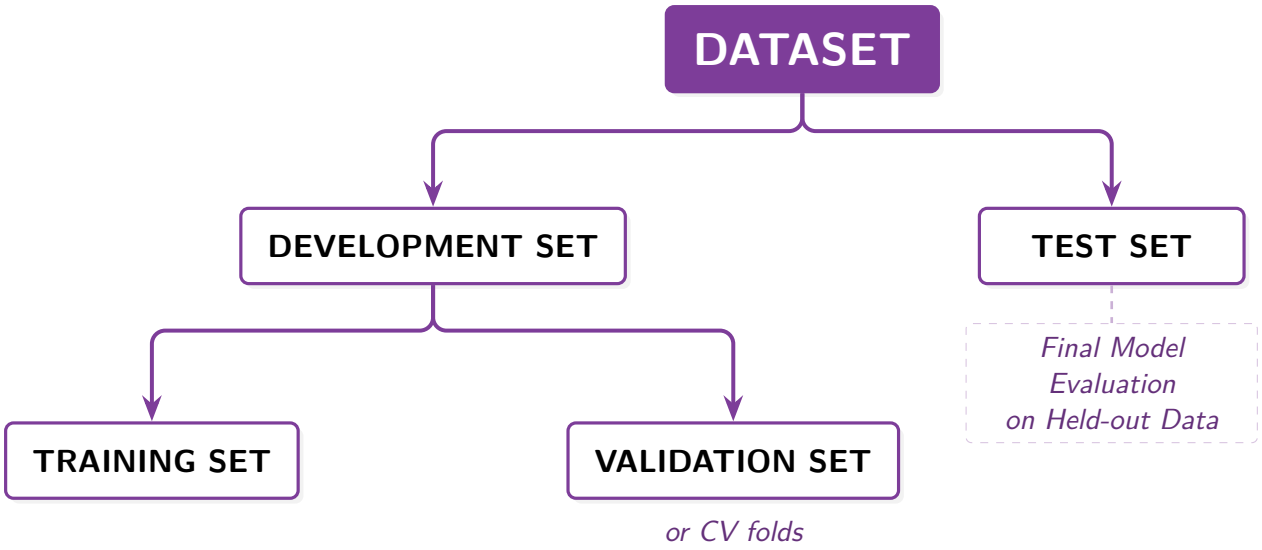


Figure 4.1: Dataset splits in modeling.

Table 4.2: Summary of utilized datasets and the rationale for test set selection.

Data Set	Test Set Selection Method	Justification
DS1	Repeated random sampling (5 iterations)	A robust protocol designed to mitigate the variance associated with a single random split.
DS2	Repeated random sampling (4 iterations)	A robust protocol designed to mitigate the variance associated with a single random split.
DS3	Repeated random sampling (4 iterations) of 1% of ILs SMILES	Ensures the model is evaluated on challenging, out-of-distribution ILs not included in the training set.
DS4	Repeated random sampling (5 iterations) (SMILES based)	A robust protocol designed to mitigate the variance associated with a single random split.
DS5	Standard random split	Protocol aligns with the methodology reported in the source data [105].
DS6	Non-overlapping paired selection (ensuring IL-solute pairs are mutually exclusive)	Protocol aligns with the methodology reported in the source data [116].

tasks were ideally selected randomly from the pool of tasks possessing sufficient data for both adaptation and evaluation. However, the available task pool may be constrained; this limitation is evident in DS 4, which comprises only four primary tasks. Consequently, the leave-one-task-out validation protocol, as depicted in Figure 4.2, was employed in this instance.

To assess the performance of the models, several metrics were utilized, including the coefficient of determination (R^2), root mean square error (RMSE), mean absolute error (MAE), and mean absolute percentage error (MAPE) [30]. The metrics are defined as follows in Equations (4.1)–(4.4):

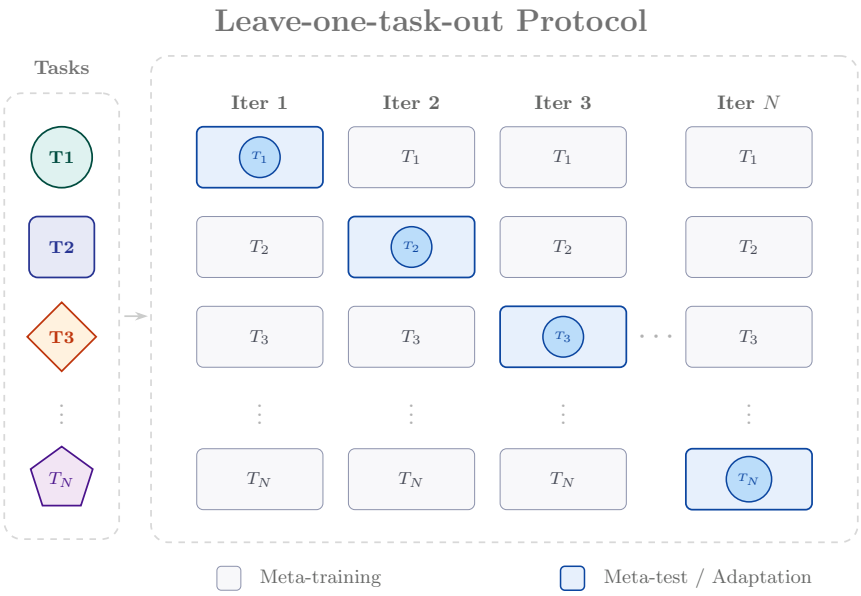


Figure 4.2: Visual explanation of leave-one-task-out protocol

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} \tag{4.1}$$

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2} \tag{4.2}$$

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i| \tag{4.3}$$

$$MAPE = \frac{100}{n}\sum_{i=1}^{n}\left|\frac{y_i - \hat{y}_i}{y_i}\right| \tag{4.4}$$

where: n is the number of observations, y_i are the actual values, $\hat{y}_i$ are the predicted values, $\bar{y}$ is the mean of the actual values.

Interpretation of the models' rationale served as an additional validation protocol. Since this study concentrates on chemoinformatics' methods rather than structure-property relationships, the models' interpretation is serving an auxiliary role. For that purpose tools like Shapley Additive Explanations (SHAP), Local Interpretable Model-agnostic Explanations (LIME), and ELI5 were utilized. The SHAP method is based on game theory concepts of contributions of inputs to increasing or decreasing of predicted output [145]. In contrast, LIME utilizes highly interpretable surrogates on local parts of inputs values ranges [146]. The analysis of decision paths within the decision trees are referenced as ELI5 method [147].

4.3 Molecular representations

Featurization of input plays an important role in the modeling workflow. However, molecules can be described numerically in various ways. Predominantly, MD and MF were utilized for classical ML.

The accurate processing of chemical data necessitates the calculation of reliable molecular descriptors and, occasionally, the optimization of molecular geometries. In this study, SMILES strings were transformed into molecular objects utilizing RDKit. Subsequent structural optimization, when required, was performed via Open Babel [148] or RDKit [142], employing a conformational search followed by force-field minimization. This methodology provides a computationally efficient and consistent alternative to more resource-intensive approaches based on quantum chemistry principles.

As mentioned in Chapter 1, MD are derived from a structure of a molecule to encode its physicochemical properties. They may include properties such as number of hydrogen bond donors/acceptors, partition coefficient, or electrostatic potential maps. MD for DS1 were utilized as calculated in its original article with Dragon software (including descriptors based on 3D structure of molecules). MD for DS2 and DS5 were calculated utilizing 2D set of descriptors implemented in RDKit package.

As defined in Chapter 1, MF are binary bit vectors. They code presence of specific substructures (subgraphs) within a molecule with value 1 and its absence with 0. Notably, Morgan fingerprints capture moieties in a circular manner around each atom. Adjusting the radius allows for capturing larger or smaller subgraphs. MF were utilized, implemented as Morgan fingerprints with radius of 2, to study DS1, DS2, and DS4.

The molecular representation learning paradigm within DL is fundamentally driven by the direct processing of raw molecular graphs or SMILES strings. Deep NN architectures typically comprise feature extraction layers followed by a regressor head. In the context of GNN, the capture of both local and global molecular features is achieved through multiple graph convolutions, which aggregate information across neighboring nodes. Initial featurization strategies for bonds incorporated attributes such as bond order and aromaticity, while nodes were characterized by atomic number, hydrogen count, hybridization, aromaticity, and charge [149]. Various charge models were evaluated, including Gasteiger-Marsili, MMFF94, formal, and QTPIE charges, each presenting distinct strengths and limitations contingent upon the molecular system and application. The readout phase involves employing a pooling function to distill a graph-level representation from the individual node embeddings. GNN models were specifically employed for the analysis of the DS3 and DS6 datasets. Conversely, transformer-based NN architectures process SMILES as a sequence of tokens, utilizing a dedicated CLS token to aggregate information from the entire sequence and generate a global representation, a methodology applied to the DS5 dataset. The regressor head is frequently implemented as a simple MLP that acts upon the derived molecular representations, such as the graph vector obtained post-readout or the CLS token embedding [150].

Overview of the utilized molecular representations is shown in Figure 4.3.

Multi-dimensional datasets were visualized using several techniques. The FreeViz [151] method facilitates the generation of visualizations characterized by robust class separability, a feature that is particularly critical for exploratory data analysis. This technique

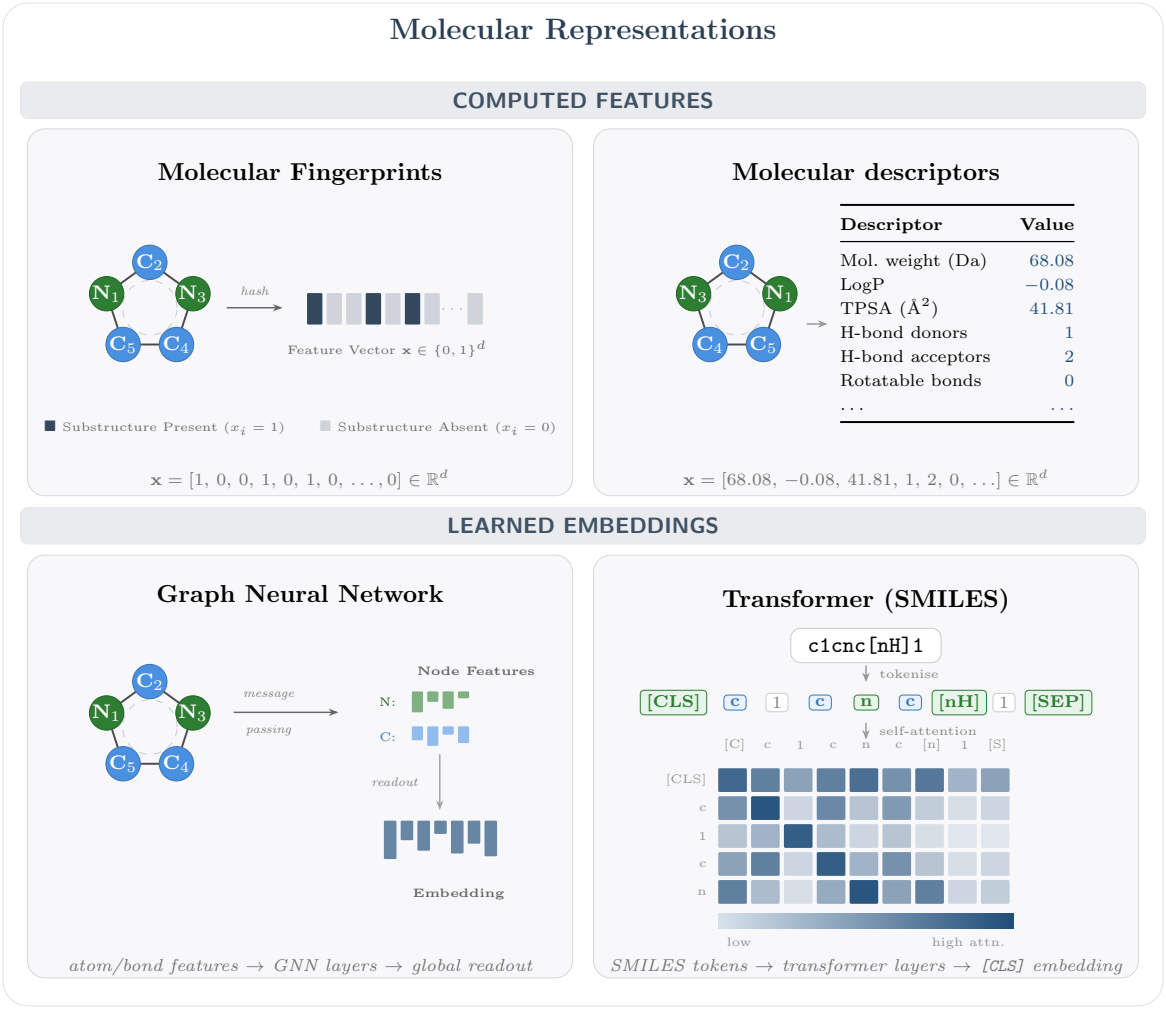


Figure 4.3: Overview of the utilized molecular representations

enables the determination of an optimal two-dimensional projection designed to effectively differentiate between samples within a dataset. The underlying algorithmic rationale can be conceptualized through a physical analogy: the method employs a mathematical construct analogous to a physical potential function, which governs the attraction and repulsion between data points. Specifically, instances exhibiting similar characteristics are attracted to one another, while those with substantial variance repel neighboring points, mirroring the physical forces between particles. Consequently, an optimal projection is identified, manifesting effective class separation within the resulting scatterplot. Within this metaphorical framework, the optimization objective—achieving maximal class separability—is conceptually equivalent to identifying the configuration (projection) that minimizes the system's potential energy.

To facilitate the visualization of the high-dimensional chemical space inherent in the dataset, the t-distributed Stochastic Neighbor Embedding (t-SNE) [152] algorithm was employed. As a non-linear dimensionality reduction technique, t-SNE projects the data into a lower-dimensional manifold while preserving the local structure and ensuring the discernible representation of inherent clusters. The resulting t-SNE visualizations elucidated the molecular relationships within the dataset, thereby identifying distinct clusters

and underlying patterns that were obscured within the original high-dimensional feature space.

The comprehension of chemical space is frequently augmented through the analysis of compound similarity. Various methodologies can be employed to quantify this similarity. Descriptor-based techniques leverage physicochemical properties (e.g., molecular weight or polar surface area) and apply distance metrics within a multidimensional feature space. More recently, machine learning-derived embeddings, particularly those generated by neural networks, have facilitated the comparison of compounds within learned latent spaces that offer a more accurate representation of bioactivity or functional similarity. A foundational approach involves the utilization of MF, which can be compared using similarity metrics such as the Tanimoto coefficient. Specifically, Tanimoto coefficients were computed based on Morgan fingerprints with a radius of 2 and a vector size of 2048 bits. The generation of these fingerprints and the subsequent coefficient calculations were performed using appropriate RDKit [142] functionalities.

The conceptualization of compound similarity within meta-learning paradigms presents a significant complexity. Specifically, the methodology for quantifying similarity must account for the inherent multi-task nature of the underlying datasets [153]. This necessitates a nuanced approach to similarity measurement, wherein the comparison between two data instances can be categorized either as within-task similarity—evaluating the relatedness of data points within the same task—or between-task similarity—assessing the relationship between data points originating from distinct tasks. The Tanimoto similarity metric was calculated across all pairwise combinations of the data points. These pairs were systematically classified into two distinct categories: intra-task pairs, representing data points belonging to the same task, and inter-task pairs, representing data points from different tasks. The resulting similarity scores derived from these pairwise comparisons were subsequently aggregated through a mean calculation to yield a comprehensive, averaged similarity measure. This conceptual framework of within-task and between-task similarity is illustrated in Figure 4.4, where task identity is represented by color and chemical space proximity by directional arrows.

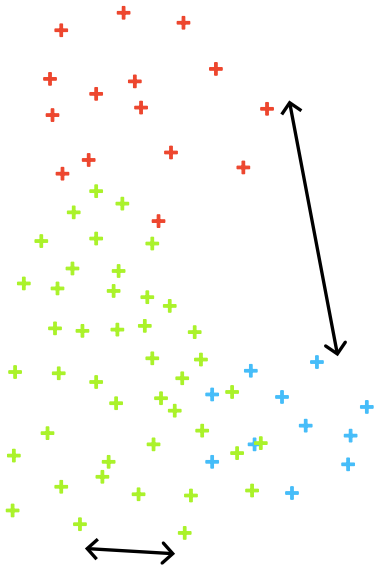


Figure 4.4: Illustration of within-task and between-task similarity concept.

4.4 Classical machine learning modeling

The feature preprocessing pipeline for the classical ML methods involved the application of standard protocols, beginning with the removal of redundant and invariant features. Subsequently, numerical features were standardized to ensure scale invariance across variables while preserving their intrinsic relative distances. To optimize predictive accuracy, a rigorous feature selection process was implemented. This selection was achieved by comparing two distinct methodologies: a ranking system derived from a gradient boosting model or the mutual information criterion. The final choice of feature subset was determined by the method that demonstrated superior performance metrics on the validation set. A suite of classical ML methodologies was employed, encompassing linear regression (MLR), ensemble decision tree approaches (RF, GB), and basic neural network architectures (MLP).

The MLR formulation, as specified in Equation 4.5, estimates the parameter vector $\hat{\boldsymbol{\beta}}$ through the minimization of a loss function quantifying the discrepancy between observed target values and those predicted by the linear model. To mitigate issues arising from high-dimensional feature spaces and multicollinearity, the model may be augmented with regularization terms, incorporating either the L1 norm (Lasso) or L2 norm (Ridge) of $\boldsymbol{\beta}$. Such regularization imposes constraints on parameter magnitudes, thereby enhancing model stability and generalization performance [154].

$$\hat{\boldsymbol{\beta}} = \arg\min_{\boldsymbol{\beta}} \left\{ \frac{1}{2n} \sum_{i=1}^{n} \left(y_i - \mathbf{x}_i^\top \boldsymbol{\beta} \right)^2 + \lambda \left(\alpha \|\boldsymbol{\beta}\|_1 + \frac{1-\alpha}{2} \|\boldsymbol{\beta}\|_2^2 \right) \right\} \tag{4.5}$$

The RF algorithm [25] derives its final predictive output through an ensemble structure of decision trees, as formally defined in Equation 4.6. This aggregation mechanism involves calculating the arithmetic mean of the predictions generated by B individual trees, denoted as $T_b(x)$.

$$F(x) = \frac{1}{B} \sum_{b=1}^{B} T_b(x) \tag{4.6}$$

While both RF and GB utilize decision trees as fundamental regressors or classifiers, their ensemble methodologies exhibit a profound architectural divergence. Unlike the parallel training paradigm of RF, GB [155] adopts a sequential construction process, wherein subsequent models are trained specifically to mitigate the residual errors generated by their predecessors. This iterative procedure defines an ensemble comprising M trees, as formally delineated in Equation 4.7.

$$F_M(x) = f_0(x) + \sum_{m=1}^{M} \eta \cdot h_m(x) \tag{4.7}$$

The MLP algorithm is characterized by a parametric network architecture comprising multiple layers of interconnected neurons. The initial layer processes the input vector x through a linear transformation, defined by the application of weights W_1 and bias b_1,

followed by a non-linear activation function σ_1. The resulting output from this layer serves as the input for subsequent layers, undergoing iterative transformations. Specifically, a basic two-layer MLP can be formally represented by Equation 4.8 [156].

$$y = \sigma_2 \left(W_2 \cdot \sigma_1 (W_1 x + b_1) + b_2 \right) \tag{4.8}$$

The optimization of the parameters θ, which encompass both weights and biases, is conducted according to Equation 4.9 with the objective of minimizing the loss function $\mathcal{L}$. This optimization procedure is implemented through back-propagation, which updates θ in proportion to the gradient $\partial\mathcal{L}/\partial\theta$ scaled by the learning rate α [157]. In the context of this study, a sophisticated variant of stochastic gradient descent, specifically the Adam optimizer, is utilized [158].

$$\theta^{t+1} = \theta^t - \alpha \frac{\partial \mathcal{L}\left(X, \theta^t\right)}{\partial \theta} \tag{4.9}$$

where $\theta = \{W^{(l)}, b^{(l)}\}_{l=1}^{L}$

4.5 Deep learning and meta-learning modeling

Deep NN architectures are fundamentally rooted in the principles of MLP models, yet they achieve enhanced representational capacity through the integration of feature extraction layers. This mechanism facilitates the automatic learning of hierarchical representations, wherein intermediate network layers progressively transform raw input data into increasingly abstract and semantically rich features. Crucially, unlike traditional, manually engineered descriptors, these learned representations inherently capture complex structural patterns, relational dependencies, and latent properties pertinent to the specific predictive objective. Consequently, these specialized layers empower deep models to process raw data modalities-such as molecular graphs or SMILES strings-directly, thereby obviating the necessity for predefined representations like MD or MF [159].

A Transformer-based NN employs an attention mechanism (Equation 4.10) [160] to encode each token - i.e., an element of the input sequence - in relation to the entire sequence context. Within this framework, SMILES representations are treated as sequences of tokens, where each token corresponds to a component of the SMILES string (e.g., atoms or bonds). Following the application of attention, the value vector V is computed using weights derived from the similarity between the query Q and the keys K. Transformer models may adopt encoder-only, decoder-only, or encoder–decoder architectures, and are typically pre-trained on large-scale datasets. A widely used example in chemoinformatics is ChemBERTa [161], which is based on the Bidirectional encoder representations from transformers (BERT) architecture [150].

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right) V \tag{4.10}$$

The integration of graph priors frequently yields enhanced performance in molecular modeling tasks, which has led to the widespread adoption of GNN within MPP research. Fundamentally, the GNN [162] architecture processes data structured as a graph G, defined by a set of vertices (nodes) V and edges E, as detailed in Equation 4.11. The operational mechanism of GNN involves iteratively updating node representations through a message-passing update function (Equation 4.12). Specifically, the graph convolution operation typically employs a learnable weight matrix W^{l+1} during the training phase (Equation 4.13) [163].

$$G = \{V, E\} \tag{4.11}$$

$$x_v^{l+1} = f_\theta\Big(x_v^l, \{\, x_u^l : u \in N(v) \,\}\Big) \tag{4.12}$$

$$x_v^{l+1} = W^{l+1} \sum_{u \in N(v) \cup \{v\}} \frac{1}{c_{u,v}} \, x_u^l \tag{4.13}$$

Training with GNN is visually shown in Figure 4.5.

The MAML framework presents a compelling methodology for the pre-training of NN. Its inherent model-agnostic nature allows for its application across diverse network architectures. The core objective of MAML is to determine the optimal parameters θ that minimize the aggregate loss $\mathcal{L}$ across the distribution of tasks $p(\mathcal{T})$, specifically considering individual tasks $\mathcal{T}_i$ (Equation 4.14). The parameter updates are governed by the optimization procedure detailed in Equation 4.15. However, this optimization process necessitates the computation of two distinct gradients, which introduces a significant computational overhead that ultimately constrains the scalability of the MAML algorithm [124].

$$\min_\theta \sum_{\mathcal{T}_i \sim p(\mathcal{T})} \mathcal{L}_{\mathcal{T}_i}(f_{\theta_i'}) \tag{4.14}$$

$$\theta \leftarrow \theta - \beta \nabla_\theta \sum_{\mathcal{T}_i \sim p(\mathcal{T})} \mathcal{L}_{\mathcal{T}_i}(f_{\theta - \alpha \nabla_\theta \mathcal{L}_{\mathcal{T}_i}(f_\theta)}) \tag{4.15}$$

Overview of modeling with MAML algorithm is depicted visually in Figure 4.6 and in pseudocode in Algorithm 1.

First-order MAML circumvents the computational complexity associated with second-order derivatives by approximating these values as zero [124]. The Reptile method further streamlines this meta-learning paradigm. It operates by sampling a specific task and performing a limited number ($k > 1$) of gradient descent steps using the associated data to update the parameters of the NN. The algorithmic formulation of Reptile is detailed in Algorithm 2 [125].

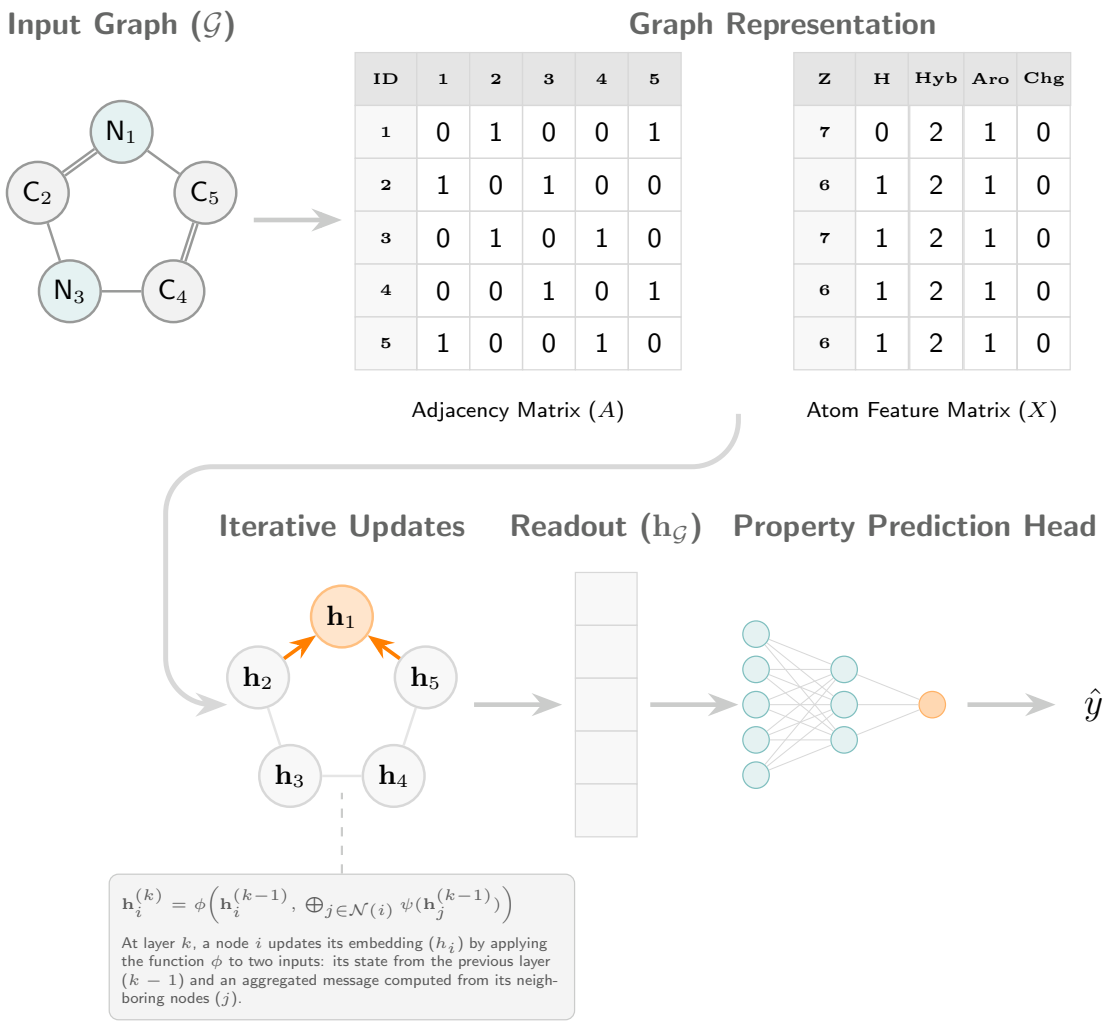


Figure 4.5: Overview of modeling with GNN algorithm

Algorithm 1 The MAML algorithm

Require: Task distribution $p(\mathcal{T})$, inner LR α, meta LR β

Initialize parameters θ
while not converged **do**
  Sample tasks $\{\mathcal{T}_i\} \sim p(\mathcal{T})$
  for each task $\mathcal{T}_i$ **do**
    Sample support set S_i
    $\theta_i \leftarrow \theta - \alpha \nabla_\theta \mathcal{L}_{\mathcal{T}_i}(\theta; S_i)$
  end for
  Sample query sets Q_i
  $\theta \leftarrow \theta - \beta \nabla_\theta \sum_i \mathcal{L}_{\mathcal{T}_i}(\theta_i; Q_i)$
end while

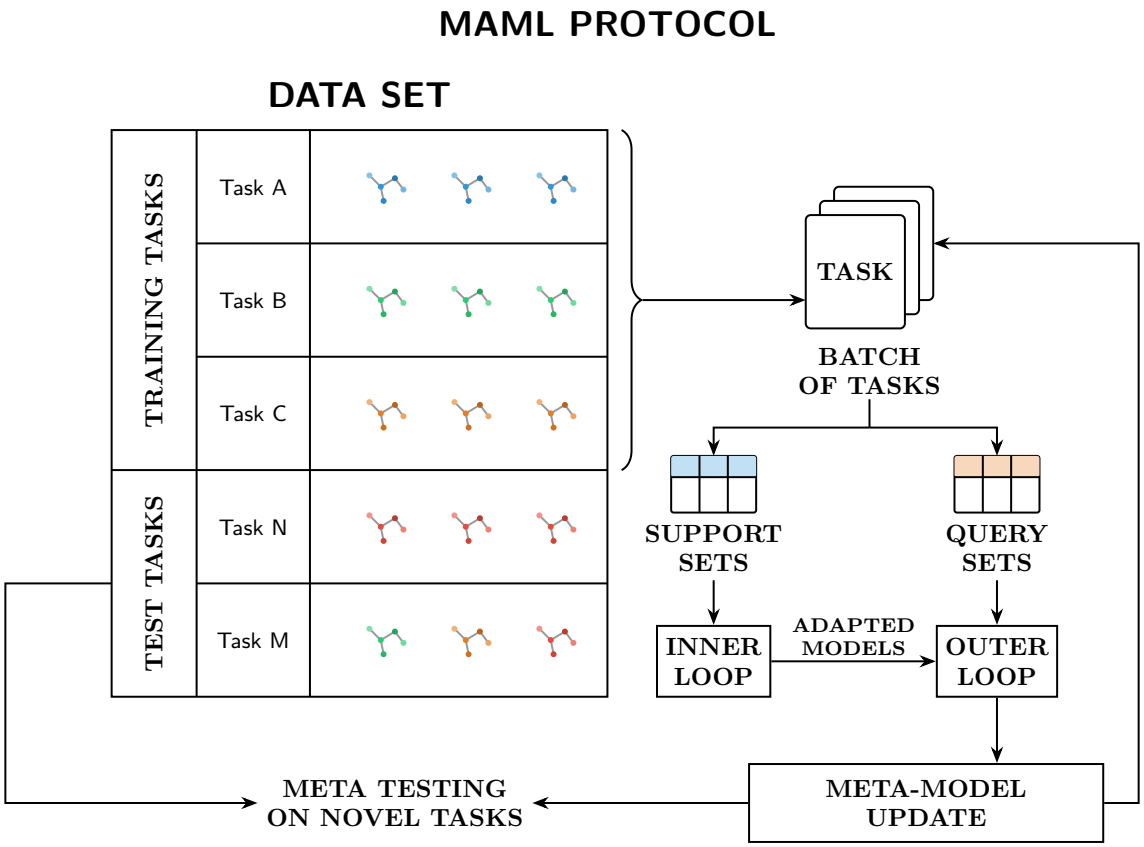


Figure 4.6: Overview of modeling with MAML algorithm

Algorithm 2 The Reptile algorithm

Require: Task distribution $p(\mathcal{T})$, LR α

Initialize parameters θ
while not converged **do**
  Sample task $\mathcal{T} \sim p(\mathcal{T})$
  for $k = 1$ to K **do**
    Sample data batch B_k from $\mathcal{T}$
    $\theta \leftarrow \theta - \alpha \nabla_\theta \mathcal{L}_\mathcal{T}(\theta; B_k)$
  end for
end while

4.6 Implementation details

In addressing the development of an effective machine learning pipeline, a series of tools and libraries were selected based on their compatibility with Python programming language and proficiency in handling various data types. The primary platform employed throughout this project was Python [164], which leverages PyTorch for neural network construction along with its related libraries such as PyTorch Geometric for GNN [165]. Additionally, Orange Data Mining [144] tools were incorporated to extract CN2 rules and implement some of the classical machine learning models.

In order to facilitate seamless data handling throughout the pipeline, Pandas [166] has been employed as a primary data manipulation tool. Scikit-learn [167], another essential Python library, was used for data preprocessing and pipelines along with the execution of classical ML algorithms. In addition to PyTorch, XGBoost [168] and LightGBM [169] libraries were incorporated into the pipeline to explore additional classical ML models.

Handling chemical data requires specialized tools; thus, RDKit [142], OpenBabel [148], and Mordred [170] have been integrated into the project, allowing for efficient preprocessing and manipulation of complex molecular structures. To gain a deeper understanding of the models under development, interpretation techniques such as SHAP [145], LIME [146], and ELI5 [147] were employed.

The evaluation of model performance is crucial in any machine learning pipeline; therefore, Permetrics library [171] was utilized to provide comprehensive metrics for assessing both classical and deep learning models. Data visualization purposes were met through the incorporation of Matplotlib [172] and Seaborn [173], the open-source plotting libraries that enable effective presentation of findings. Furthermore, meta-learning techniques have been explored using deepchem [174] and learn2learn [175] libraries.

Chapter 5

Summary of results and discussions

5.1 Classical low-data property prediction

The study commenced with utilization of classical ML methods. The datasets under investigation represented current topics in chemistry of multimolecular systems, particularly carbon dioxide absorption with ILs (DS1) and lignin isolation from biomass with ILs (DS2).

The initial phase of the DS1 modeling involved an evaluation of applicability of Fractional Free Volume (FFV) as a molecular descriptor. The utility of FFV is acknowledged, yet its scalability to large-scale datasets presents a significant computational challenge. Consequently, the feasibility of rapidly approximating FFV using alternative metrics (that possess lower computational overhead than traditional molecular dynamics simulations) was investigated. To empirically validate these proxies, the Pearson correlation coefficient was calculated between FFV and several descriptors: the solvent accessible surface area (SASA), FFV estimated via energy minimization within a small cluster (approximately 20 ion pairs), and FFV derived directly from the force-field minimized structure of a single ion pair. The comparative analysis revealed that SASA exhibited the strongest correlation with FFV ($r = 0.62$). However, the correlation coefficients for the small cluster estimation ($r = 0.58$) and the single-molecule estimation ($r = 0.45$) were substantially lower. Despite the superior performance of SASA, the correlation was insufficient to justify its use as a direct substitute for FFV.

The prediction of the FFV value from MD has been demonstrated to be challenging when employing linear models. Consequently, advanced ML methodologies were utilized to address this inherent difficulty. The incorporation of ARKA descriptors facilitated the development of a predictive model for the FFV value, which achieved an R^2 of 0.779 on the test set. While this model exhibits superior performance compared to previously reported approaches, its overall predictive power remains suboptimal as some information on FFV is not available in MD. The results of the modeling are shown in Table 5.2 (adapted from publication A3 of the series).

The performance of ML methods could be further enhanced with MD crafted particularly for sorption modeling. Even though preliminary studies have shown that inclusion of

Table 5.1: Correlation of FFV proxies with reference FFV.

descriptor	calculation method	Pearson correlation (R) to FFV
FFV	energy minimization followed by molecular dynamics using a large cluster of molecules	reference
SASA	energy minimization using a single molecule	0.62
FFV estimation	energy minimization using a small cluster of molecules	0.58
FFV estimation	energy minimization using a single molecule	0.45

Table 5.2: Comparison in performance between MLP and GB models while trained for predicting the FFV value.

modeling approach	CV R^2	CV $RMSE$	test R^2	test $RMSE$
GB	0.474	0.016	0.700	0.013
MLP	0.605	0.014	0.779	0.011

FFV improved performance of MLR models, its' impact on more complex ML models warrants further evaluation. For this purpose, GB and MLP algorithms were utilized to train models for predicting Henry's Law Constant both with and without FFV among molecular descriptors. Table 5.3 (adapted from publication A3 of the series) shows how addition of FFV changes performance of ML models.

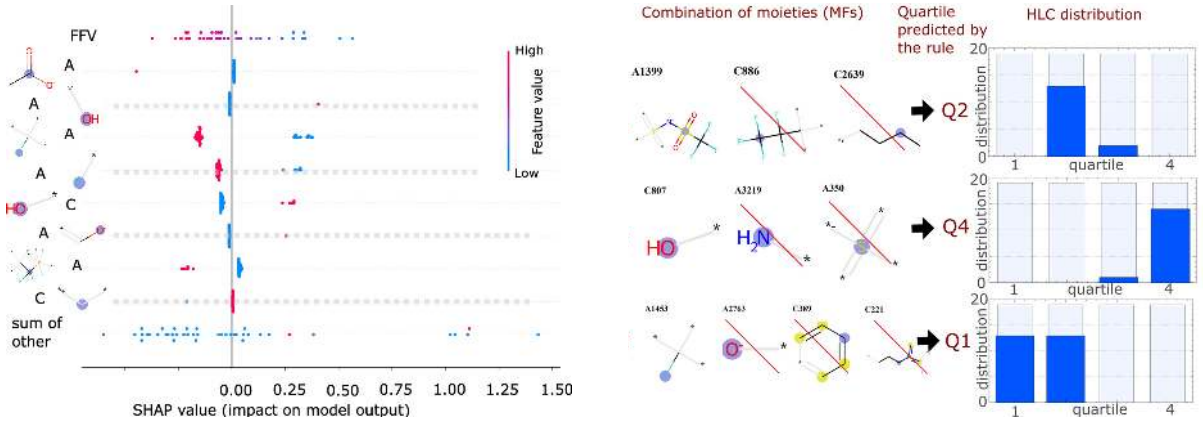
The MLP model demonstrated substantially superior performance compared to the GB models, yielding markedly better results across both CV and test metrics. Furthermore, the efficacy of the MLP model was enhanced by the incorporation of the FFV descriptor. The models that incorporated the FFV descriptor demonstrated superior performance relative to the baseline model utilizing only molecular descriptors. The integration of the FFV descriptor into the MLP model resulted in enhanced performance across multiple metrics. Specifically, the descriptor improved cross-validation performance, as indicated by the increase in the CV R^2 value from 0.712 to 0.775. Moreover, the observed decrease in the $RMSE$ (from 1671 to 1710 for MLP utilizing FFV) substantiates the hypothesis that the FFV descriptor effectively reduced the magnitude of prediction errors. The validation on a test set revealed that the R^2 value achieved by the MLP model on the test set increased substantially, rising from 0.738 to 0.833. This finding underscores the efficacy of employing nonlinear ML methodologies in significantly enhancing model predictive accuracy.

The developed ML models facilitate the extraction of chemical insights regarding carbon dioxide solubility in ILs, as illustrated in Figure 5.1 (adapted from publication A3 of the series). The SHAP analysis (Figure 5.1a) identifies free volume (FFV) as the most influential descriptor. Specifically, a positive correlation was observed between increased free volume and enhanced solubility, suggesting that the structural voids within ILs facilitate the initiation of CO_2 absorption. Furthermore, the Shapley value analysis provides critical insights into the specific chemical moieties that govern the predicted sol-

Table 5.3: Comparison in performance between MLP and GB models while trained with or without FFV descriptor.

modeling approach	CV R^2	CV $RMSE$	test R^2	test $RMSE$
MLR	0.385	2378	0.711	2464
MLR with FFV	0.466	2346	0.730	2379
GB	0.646	2178	0.507	2926
GB with FFV	0.641	2224	0.617	2580
MLP	0.712	1671	0.738	2386
MLP with FFV	0.775	1710	0.833	1859

ubility. The presence of hydroxyl groups in both anions and cations was identified as a significant factor promoting high values of the Henry's Law Constant. Additionally, the SHAP diagram linked two distinct fingerprints—associated with different chemical environments of a fluoride atom—to predictions of elevated carbon dioxide solubility. The CN2 analysis (Figure 5.1b) further enables the investigation of the impact of combined chemical moieties. CO_2 sorption in ILs is fundamentally governed by two primary mechanisms: specific molecular interactions and free volume effects. While fluorinated anions enhance CO_2 uptake through favorable interactions with fluorine atoms, often resulting in moderate-to-high solubility, rule-based models such as CN2 may inadequately represent these effects by excluding chemically relevant features in pursuit of statistical confidence. Hydroxyl-only ILs (lacking amines or fluorinated anions) exhibit high HLC, indicative of weak specific interactions. Conversely, when anion fluorination is present, it tends to dominate the sorption behavior, thereby diminishing the relative importance of other structural features, such as aromatic rings or long alkyl chains.



(a) SHAP diagram (b) CN2 rules

Figure 5.1: Chemical insights on carbon dioxide solubility in ILs.

Classical ML methodologies were also employed to investigate the DS2 dataset. The primary objective of the study was to address an existing gap in the literature by developing a predictive framework for lignin extraction capacity from biomass utilizing ILs as solvents, as shown in Figure 5.2 (adapted from publication A3 of the series). The core of the research lies in the integration of three distinct categories of input data into a unified predictive model capable of accurately forecasting extraction yield. The process is influenced by biomass composition, process parameters (like temperature and time), and ILs

structural characteristics. Given that the extraction process is inherently multivariate, its outcome is simultaneously dependent on all three input domains. These factors can contribute independently; for instance, certain solvents may exhibit optimal performance only within specific temperature ranges, or biomass compositions may necessitate varying degrees of hydrogen-bond basicity in the solvent to achieve efficient extraction. Similarly to DS1, the dataset comprised a limited sample size, consisting of only about 110 records. The dataset focused on herbaceous biomass, which was selected over woody biomass data due to its superior volume. However, the integration of both subsets into a single model proved infeasible, as visual inspection via FreeViz revealed a distinct separation between the two subgroups.

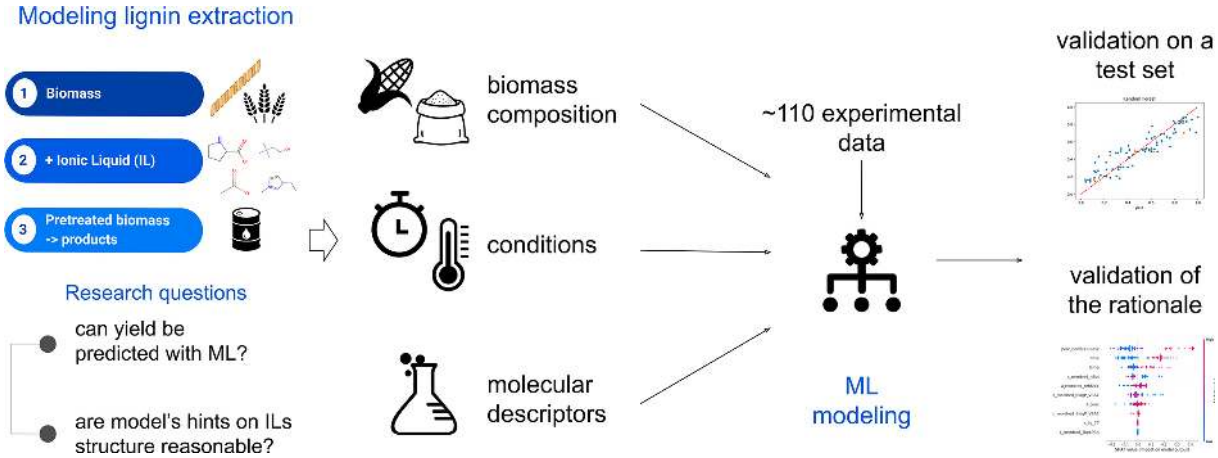


Figure 5.2: Study design on lignin isolation yield prediction

The predictive models were constructed utilizing GB and RF algorithms. These decision tree-based methodologies were selected for modeling the process due to their demonstrated efficacy with limited datasets. The molecular representations employed included either MD or MF. A significant challenge arose because certain features within MD and MF lacked interpretability or direct correlation with the lignin isolation yield. When the descriptor set was unrestricted, many features were inherently difficult for human interpretation. The full model, which incorporated complex descriptors such as AATSC3i (autocorrelation of lag 3 weighted by ionization potential), MATS3pe (Moran coefficient of lag 3 weighted by pauling electronegativity), AXp-5d (5-ordered averaged Chi path weighted by sigma electrons), and MDEO-11 (molecular distance edge between primary O and primary O), was consequently classified as a ”blackbox” model. To enhance model transparency and interpretability, a second model was developed using only a pre-selected subset of descriptors, which was designated as the ”interpretable” model. The selected descriptors included aromatic atoms count, number of hydrogen bond donor / acceptors, estimated partition coefficient, rotatable bonds count, topological polar surface area, and molecular weight. The performance metrics for these models in predicting lignin isolation yield are presented in Table 5.4 (adapted from publication A2 of the series).

The models were subjected to a rigorous inspection of their underlying rationale. This interpretability analysis was conducted utilizing SHAP, LIME, and ELI5 methodologies. Analysis of the Shapley values revealed consistent regularities across all developed models: the efficiency of the lignin extraction process was predominantly governed by the biomass composition, specifically the percentage of hemicellulose, alongside two critical process parameters: time and temperature. Furthermore, the ELI5 model interpretation was

Table 5.4: The evaluation metrics of models predicting lignin isolation yield.

modeling approach	MD-based model (test set R^2)	MF-based model (test set R^2)
GB blackbox	0.77	0.77
RF blackbox	0.71	0.75
GB interpretable	0.73	0.68
RF interpretable	0.63	0.71

performed on three randomly selected data records representing low yield (record no. 65), medium yield (record no. 1), and high yield (record no. 46). This comparative analysis highlighted significant divergences between the RF and GB models. For the GB model, the MD frequently exhibited impacts comparable to those of the process parameters. This was evident in the low-yield sample, where nRot and time demonstrated highly similar importance and impact (both in magnitude and sign), and in the high-yield sample, where the impact of nHBAcc leveraged the influence of temperature.

The analysis of LIME results, Figure 5.3 (adapted from publication A2 of the series), revealed several consistent trends regarding the prediction of lignin yield. Across all three samples, the percentage of hemicellulose (perc_hemicellulose) emerged as a dominant predictive feature. Specifically, high hemicellulose content (exceeding 0.56 units) was identified as a strong positive predictor for lignin yield, whereas intermediate ranges (0.15–0.35) were associated with predicted low yields. Furthermore, thermodynamic parameters, namely temperature and time, were found to significantly influence the model predictions, although their effects exhibited non-linear characteristics. In the context of medium yield samples, reaction time served as the most significant positive factor; conversely, in low yield samples, insufficient reaction time (very short durations) was identified as a major negative contributor. Low temperatures consistently correlated with decreased predicted yields, manifesting as a negative contribution across both medium and low yield examples. A comparative analysis of the models indicated that the RF model demonstrated a significantly greater sensitivity to the impact of a limited number of features compared to the GB model. However, the specific importance assigned to individual descriptors varied between the LIME and ELI5 methodologies. For instance, in sample 46, LIME assigned a substantially higher importance to the descriptor $nRot$ than ELI5. This suggests that a low count of rotatable bonds—likely linked to the rigid aromatic structure of the cation—and indirectly implies the high relevance of $\pi - \pi$ interactions involving the aromatic imidazolium ring. Conversely, for samples 1 and 65, the absence of aromaticity in the cation structure resulted in lower target values, as evidenced by the LIME graphs for the GB model. The influence of this structural characteristic appears to be comparable to, or even more pronounced than, the impact of temperature.

The successful application of traditional ML models, prevalent in the ILs modeling domain, does not simply transfer with transition into DL applications, however.

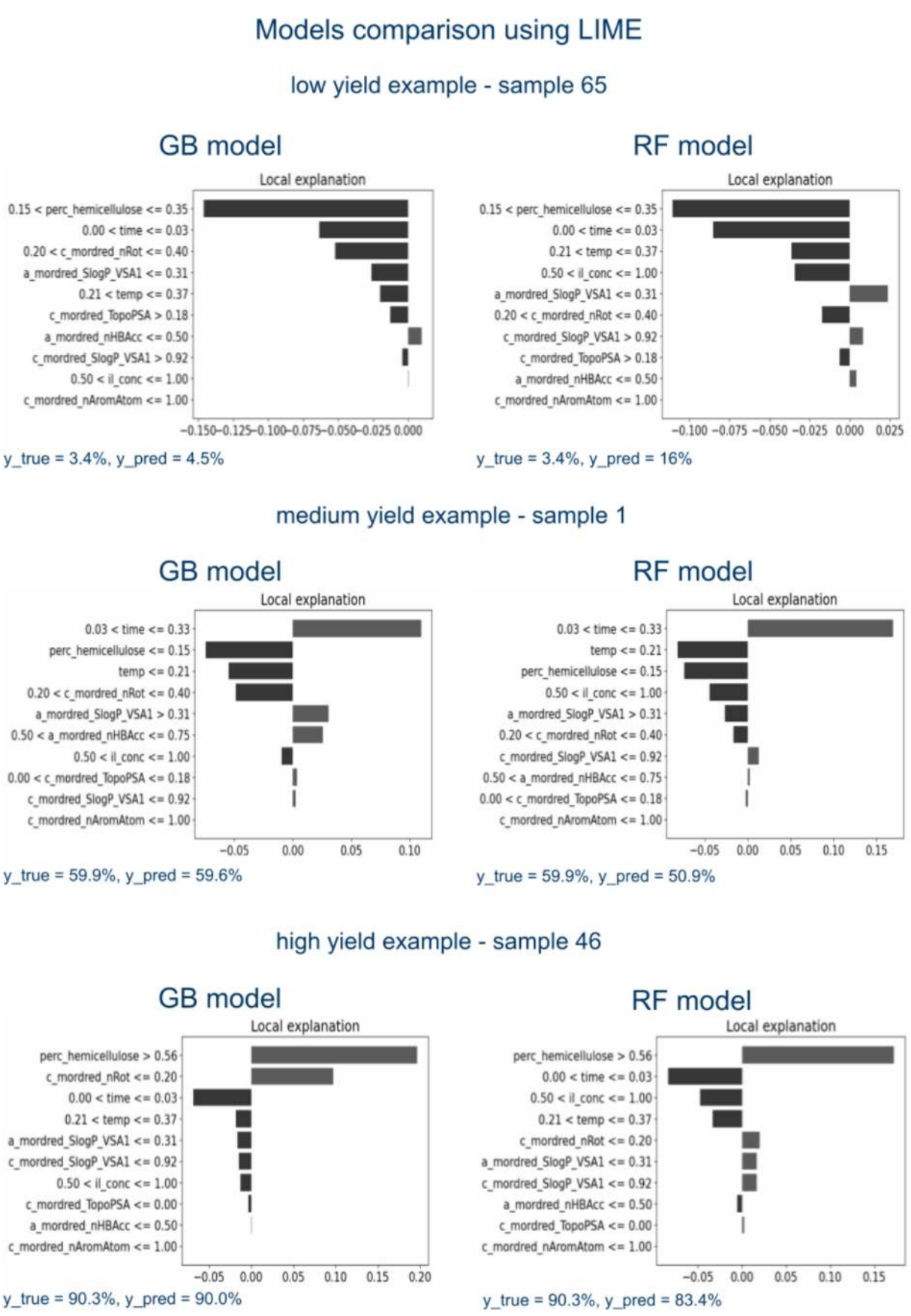


Figure 5.3: Analysis with LIME of the lignin isolation yield prediction model.

5.2 Graph Neural Networks and bimolecular systems

The study of GNN adaptation to specificity of ILs requires utilization of large datasets. Consequently, DS3 including data on density, viscosity, and surface tension was utilized. Several fundamental questions guided the study. First, how does the use of expertise-based, cleaned data—even if slightly less reliable—affect modeling outcomes? Second, what are the most effective ways to encode ionic pairs (cations and anions) into a graph structure? Third, to what extent is electrostatic information necessary for the performance of GNNs? Fourth, how do structure optimization tools (like Open Babel and RDKit) influence GNN performance? Finally, what is the potential for enhancing model performance through the application of transfer learning and fine-tuning? The problems under investigation in that context are visually summarized in Figure 5.4 (adapted from publication A1 of the series).

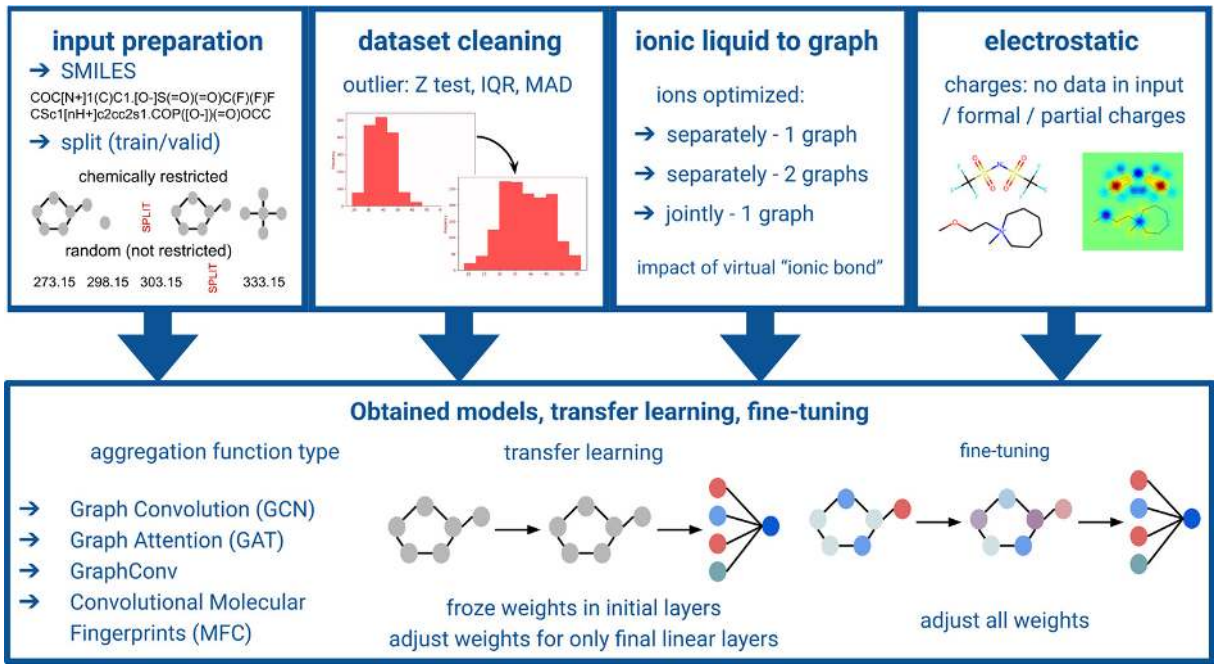


Figure 5.4: Overview of factors and methods investigated in the context of GNN applications for bimolecular systems

Data reported in the literature often exhibits inconsistencies across different systems. These discrepancies are frequently attributable to variations in measurement protocols, fluctuations in water content, and the presence of impurities, all of which can influence the target values. To ensure data comparability and homogeneity, the original benchmarking studies implemented rigorous preprocessing informed by domain-specific chemical knowledge. Consequently, the designation "clean" dataset denotes the dataset following the specific preprocessing procedures detailed in the original article. Conversely, the term "raw" dataset refers to the data as initially collected and reported in the original publications, prior to any critical assessment or refinement.

The predictive performance for density (MAD detection, $R^2 = 0.955$) and surface tension ($R^2 = 0.79$) was superior when utilizing raw datasets compared to models trained exclusively on curated, clean datasets. This finding suggests that GNN exhibit a high degree of robustness when processing mislabeled data. This phenomenon can be attributed to the structural diversity inherent in the raw dataset, which encompasses a broader chemical

space of ILs. The increased structural heterogeneity allows the GNNs to adapt more effectively to the underlying molecular information. Specifically, the exposure to a more diverse training set facilitates the optimization of convolutional layers, enabling the extraction of more generalized molecular features. Conversely, the viscosity models demonstrated a deterioration in statistical parameter performance. This contrasting outcome is likely attributable to the presence of numerous outliers within the viscosity dataset, where a subset of data points exhibits magnitudes up to seven orders of magnitude greater than the mean.

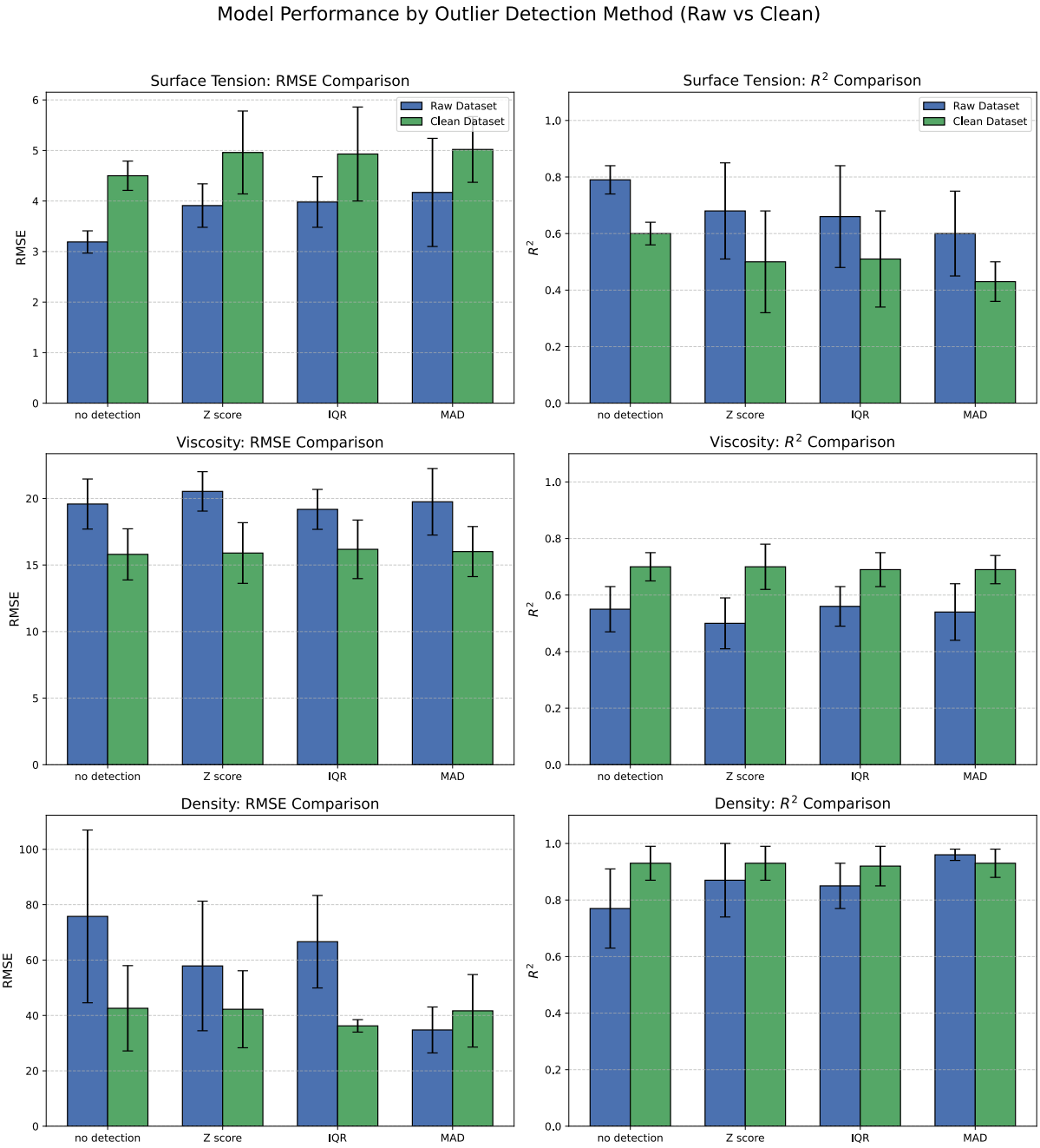


Figure 5.5: GNN models performance by outlier detection method

The modeling framework incorporated charge assignments for systems where the cation and anion were optimized through both separate and joint procedures. A total of ten configurations were evaluated, derived from five distinct charge calculation methodologies:

no charge, formal charges, Gasteiger-Marsili, MMFF94, and QTPIE partial charges. The results of these evaluations are presented in Table 5.5 (adapted from publication A1 of the series). While the utilization of formal charges among the graph input yielded marginally superior metrics, the analysis indicates that the optimization strategy employed for the ionic liquid molecule exerts a more substantial influence on the model's performance than the specific method used for charge calculation.

Table 5.5: Comparison of R^2 for separately and jointly optimized representations with different charge representations.

Charge representation	separately optimized (test set R^2)	jointly optimized (test set R^2)
No charges	0.79 ± 0.06	0.74 ± 0.11
Formal charges	0.79 ± 0.05	0.76 ± 0.07
Gasteiger charges	0.78 ± 0.04	0.74 ± 0.11
MMFF94 charges	0.78 ± 0.04	0.73 ± 0.08
QTPIE charges	0.77 ± 0.05	0.72 ± 0.09

The representation of ILs within the graph-based framework was explored through three distinct methodological approaches. In the first scenario (A), the cation and anion molecular structures were optimized independently, treating each as a separate molecular graph. These optimized graphs were subsequently integrated into a single matrix structure, where the cation graph occupied the upper-left quadrant, the anion graph was positioned in the lower-right quadrant, and the remaining matrix elements were populated with null values. Conversely, the second scenario (B) treated the cation–anion pair of ILs as a unified graph, allowing for the simultaneous optimization of the entire molecular structure. To maintain structural consistency, atoms originating from the cation were systematically listed prior to those from the anion. Finally, the third scenario (C) involved performing two parallel graph convolutions—one dedicated to the cation and one to the anion—rendering matrix concatenation inapplicable. The scenarios are shown in Figure 5.6. The analysis indicates that optimal predictive performance was achieved when ILs were modeled as a single graph, provided that the constituent ions were optimized independently. This suggests that ILs can be conceptually treated as a mixture, where the convolutional layers are capable of capturing inter-ionic interactions within the interstitial volume. Although the differences in performance across the various modeling scenarios were marginal, particularly when considering the standard deviation, it is noteworthy that the optimal configuration also yielded the lowest standard deviation. Consequently, this modeling protocol demonstrates the highest degree of reliability.

The fine performance of the aforementioned models builds a foundation for further studies. Since the GB model operates well with limited dataset (76 records), the model meta-trained on many gas-ILs systems should also work fine. Similarly, the model on lignin isolation should form a basis for subsequent modeling of isolation of hemicellulose and cellulose. They all could benefit from more deep architectures. This was shown to be limited by dataset size needed for GNN modeling, however. The aspiration to overcome this limitation raises need for utilization of meta-learning protocols.

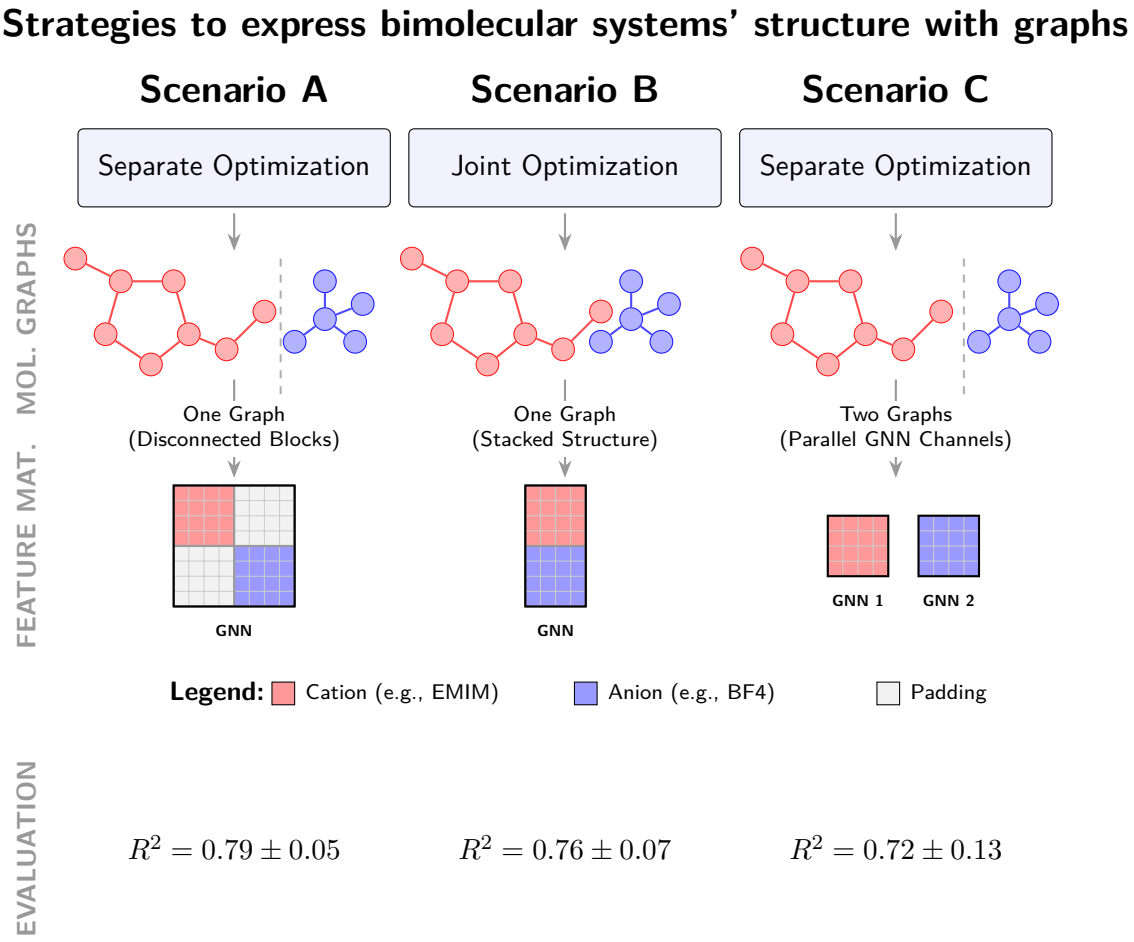


Figure 5.6: Scenarios of transitioning from bimolecular structure to graph

5.3 Preliminaries on meta-learning applications

The insights obtained from classical ML formed a basis for studies on few-shot learning. For this purpose, meta-learning was utilized. In this protocol, two-step approach utilizes meta-training on similar tasks (e.g. solutes or toxicity endpoints) and adaptation on a task of interest (test task). This method was applied in preliminary research to dataset DS1. The task was defined as a prediction of solubility of a particular solute (in the case of DS1 dataset - gases) in ILs. The meta-training set contained 84 data points in total for 15 different gases.

The scarcity of data resulted in a need for its augmentation. In total, 27209 simulated data points for 6 tasks were obtained with COSMO-RS simulations. The utilization of this data requires an appropriate training protocol, however. A simple merge of simulated and experimental data might result in conflicts between experimental and simulated structure-property relationships. Consequently, it was decided to first meta-train the MAML model with simulated data. This step was crucial for the model to encode physically plausible relationships. In the second step, the model was further meta-trained with more reliable experimental data. The two step protocol of meta-training MAML model with synthetic data is shown in Figure 5.7.

Table 5.6 (adapted from publication A3 of the series) illustrates the application of MAML, which was meta-trained according to the previously described two-step protocol. Con-

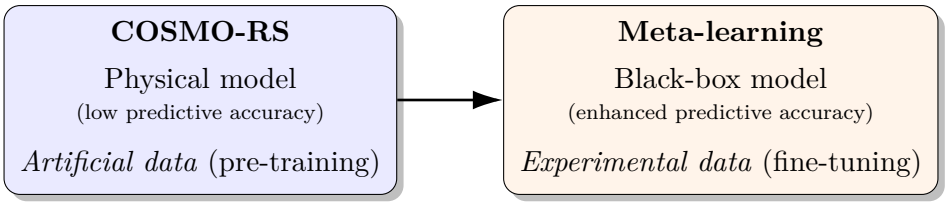


Figure 5.7: Two step protocol of training MAML model with synthetic data

sequently, all models utilized the MF training method, as the computational expense associated with 3D MD was deemed prohibitive. The efficacy of the MAML algorithm was initially assessed using experimental data, where training solely on this dataset yielded no performance enhancement and resulted in suboptimal model performance ($R^2 < 0.6$). This poor outcome is primarily attributed to the limited scale of the meta-training dataset, which caused the adaptation phase to dominate the training regimen, effectively mimicking training from scratch, thereby indicating that meta-initialization offered no advantage over a random initialization. Conversely, the incorporation of COSMO-RS predicted values into the training set significantly improved model performance. In this augmented scenario, the MAML-based NN model surpassed the performance metrics of the classical NN model. Quantitative analysis revealed an increase in the training set R^2 metric from 0.648 to 0.708. Furthermore, pretraining utilizing simulated data demonstrated a substantial performance boost, elevating the test set R^2 from 0.518 to 0.643, confirming the algorithm's capacity to effectively leverage simulated data for performance enhancement. Nevertheless, the MAML models exhibited lower performance metrics when compared to a baseline NN model augmented with FFV among MD ($R^2 = 0.643$), which achieved a best-case scenario R^2 of 0.833. To enhance the accuracy and generalizability of these models, the utilization of larger, more robust datasets should be prioritized. Moreover, expanding the scope of these models by applying MAML to alternative solubility measures, such as weight or mole fraction, which are prevalent in the literature, is warranted to broaden their utility in predictive modeling applications.

Table 5.6: Comparison in performance between MLP and MAML (augmented with simulated COSMO-RS data) for HLC prediction

modeling approach	test set R^2	test set R^2
Multilayer Perceptron (MLP)	0.629	0.636
MLP with FFV descriptor	0.648	0.670
MAML MLP-based	0.590	0.518
MAML MLP-based (two-step protocol)	0.708	0.643

Consequently, the investigation of the MAML method was warranted for larger datasets. The availability of carbon dioxide solubility data is significantly enhanced when expressed as a mole fraction of CO_2 in ILs. Figure 5.8 (adapted from publication A3 of the series) illustrates the evolution of the model metrics, specifically R^2 and RMSE, as a function of the adaptation set size. The results demonstrate that the MAML model maintained a consistent and highly stable training protocol, even when fine-tuned using only 32 samples. This observation suggests that the model was effectively trained to facilitate rapid adaptation to novel problems using limited data. Given that the meta-training dataset lacked specific data regarding the CO_2 mole fraction in ILs, this outcome is an intriguing finding, indicating that the solubility of novel gases can be successfully modeled

even with restricted data when employing MAML. When contrasted with the traditional GB model, MAML exhibits superior performance for adaptation sizes below 256 samples. However, in regimes characterized by higher data availability, the GB model demonstrates the capacity to outperform neural network approaches.

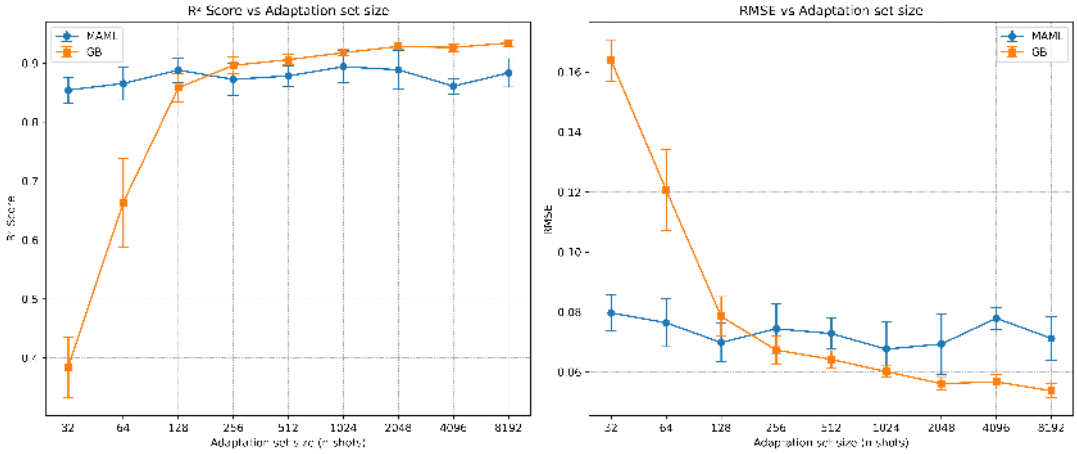


Figure 5.8: Metrics of meta-models for gas solubility prediction (adapted for carbon dioxide task).

The demonstrated efficacy of MAML when applied to a substantial dataset concerning gas absorption, characterized by mole fraction, necessitates a broader investigation into its applicability across diverse domains.

5.4 Attentive meta-learning for toxicity prediction

A promising application for MAML lies in the domain of toxicity prediction. Chemical toxicity databases, which document measured toxic effects across various biological assays, can be conceptualized as a collection of data where each specific endpoint constitutes an individual task for a meta-learning model. This framework is particularly compelling because it enables the meta-model to acquire adaptive capabilities by leveraging prior meta-training on related or analogous endpoints. However, significant modeling challenges persist, primarily due to the scarcity of toxicity data, a constraint that is exacerbated when addressing less frequently studied endpoints. This necessitates an investigation into the efficacy of MAML when applied under conditions of severely limited task availability for meta-training.

The DS5 dataset encompasses data pertaining to four primary toxicity targets: E. coli, IPC-81, AChE, and Vibrio fischeri. Prior investigations utilized the GB algorithm, predicated on the assumption that decision tree-based methodologies are effective for small sample sizes, typically ranging from 100 to 200 instances. However, the influence of dataset size on the resulting performance metrics remains insufficiently explored. This analysis aims to evaluate the efficacy of meta-learning as a viable alternative to conventional methodologies.

The benchmarking of molecular representations using the GB algorithm across four tasks revealed the absence of a universally optimal representation (Table 5.7). While 2D molecular descriptors and ChemBERTa embeddings demonstrated superior performance in tasks

such as E. coli and IPC-81, molecular fingerprints proved more effective for AChE and Vibrio fischeri. To ensure the reliability of the few-shot learning assessment, a representation capable of consistent performance across all scenarios was prioritized. Consequently, 2D molecular descriptors were selected for subsequent modeling, as they consistently achieved R^2 values exceeding 0.8 across all tasks, a threshold not met by either transformer embeddings or molecular fingerprints.

Table 5.7: Performance of GB models using 2D RDKit descriptors and ChemBERTa embeddings relative to the molecular fingerprints baseline from [105]. ΔR^2 and percentage change are computed with respect to the MF baseline.

Test task	MF baseline (R^2)	2D RDKit descriptors		ChemBERTa embeddings	
		ΔR^2	% change	ΔR^2	% change
E. coli	0.800	+0.090	+11.3%	+0.080	+10.0%
AChE	0.830	−0.029	−3.5%	−0.021	−2.5%
IPC-81	0.770	+0.038	+4.9%	+0.060	+7.8%
Vibrio fischeri	0.870	−0.030	−3.4%	−0.096	−11.0%

Figure 5.9 illustrates the evolution of performance metrics across varying adaptation set sizes. The results indicate that the traditional GB method maintains robust performance when the training set comprises at least 64 samples across the evaluated benchmarks. This observation is consistent with the established understanding that decision tree-based approaches are generally effective even with limited data; however, their predictive utility diminishes significantly when the dataset size is substantially reduced. Conversely, MAML demonstrates the capacity to effectively leverage the structure-toxicity patterns acquired during the meta-training phase.

The performance of the MAML model exhibited variability across the investigated tasks. This heterogeneity may primarily be attributed to the degree of similarity between the test tasks and the training tasks, or to the extent of chemical space overlap between the adaptation and meta-training sets. The former remains difficult to assess within the current experimental framework. Two of the endpoints, E. coli and Vibrio fischeri, are bacterial, and their performance trends may be correlated due to shared biological characteristics. In contrast, the IPC-81 utilizes rat leukemic cells, while the AChE task focuses on a specific molecular target.

The AChE toxicity data appears distinct from the other datasets, as it pertains to an enzyme rather than a whole organism or cell line. To quantitatively evaluate this distinction, the Tanimoto similarity was computed for all possible pairs of data points, categorized as either within-task (same task) or between-task (different tasks), and the resulting values were averaged. The AChE dataset demonstrated a relatively high within-task similarity of 0.244 compared to the other tasks, yet it exhibited significant dissimilarity when compared to other tasks, such as a between-task similarity of 0.132 with the E. coli task. This pronounced distinction suggests that the unique nature of the AChE dataset, characterized by both high internal cohesion and low external similarity, may be a contributing factor to the observed underperformance of the MAML model on this specific task relative to the others.

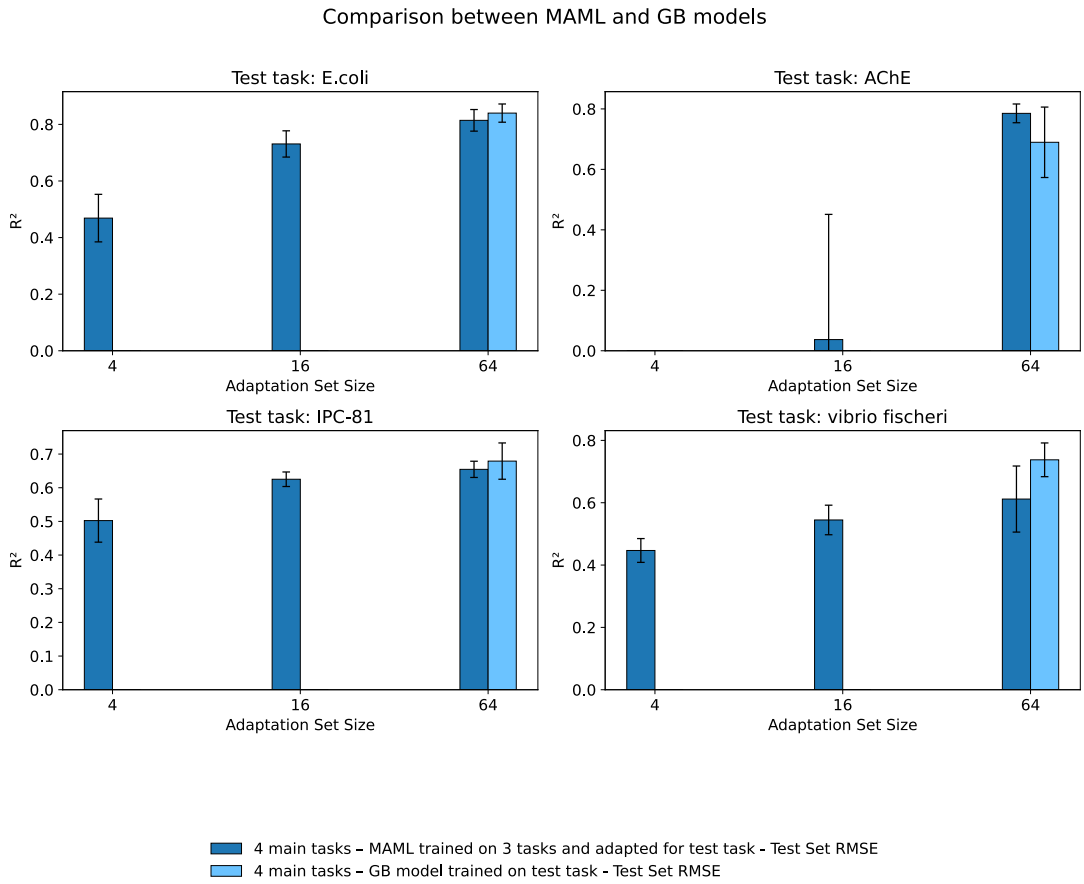


Figure 5.9: Comparison between MAML and GB models (negative metrics values not shown)

The comprehensive ILTox database encompasses a broader range of data than the subset employed in this study, yet its overall utility remains subject to critical evaluation. While the dataset is extensive, the coverage of chemical space and biological endpoints is notably uneven. Specifically, the chemical space covered is often task-specific rather than providing broad coverage across the most common ILs. Furthermore, despite the full ILTox dataset containing significantly more tasks (221), the average record count per task is approximately 8, which is substantially lower than the $\sim$ 160 records typically found in higher-fidelity subsets of the ILTox database. Figure 5.10 illustrates the performance of MAML models trained under three distinct meta-training scenarios. The baseline scenario (green) utilized four primary tasks from the ILTox database, selected for their high reliability and data availability. The second scenario (blue) employed a subset of the ILTox database curated to match the endpoint categories (animals, cells, microorganisms, plants) of a test task. The third scenario utilized the entire ILTox database. The results indicate that increasing the size of the meta-training set did not consistently lead to improved model performance. In certain instances, such as adaptation to the E. coli task, the performance differences across the scenarios were minimal. Conversely, for tasks like AChE or Vibrio fischeri, the utilization of the full database resulted in generally inferior model performance.

The random selection of less frequently investigated tasks from the ILTox database was employed to validate the predictive capabilities of the MAML model. Specifically, across

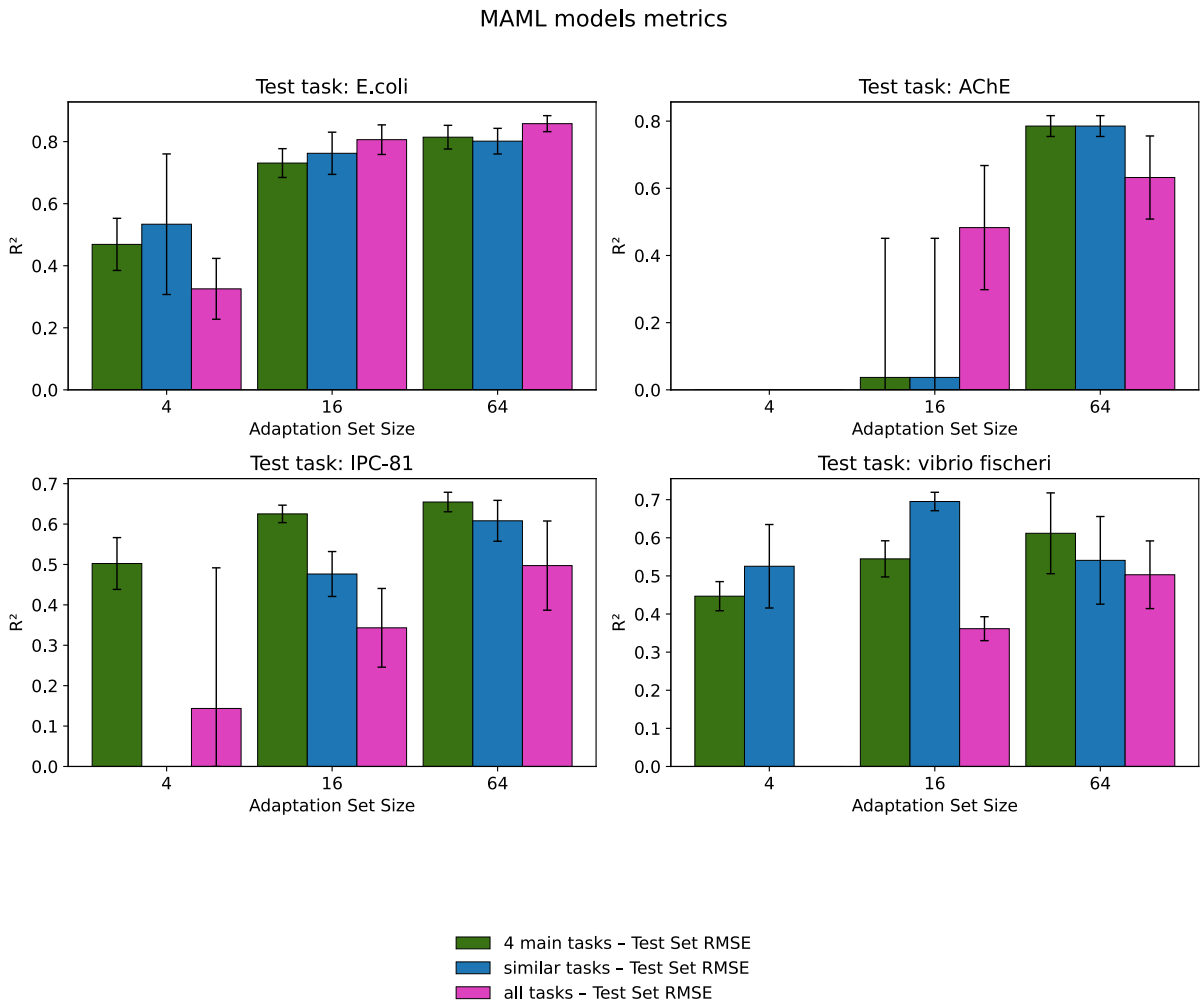


Figure 5.10: Metircs of MAML model utilizing different meta-training datasets (green - MAML trained using only 4 main tasks, blue - MAML meta-trained using tasks similar to the test task, pink - MAML pre-trained using all tasks, negative metrics values not shown)

twelve randomly selected tasks, the MAML model demonstrated qualitative performance with a MAPE below 40% on ten of the twelve tasks when utilizing an 8-shot adaptation protocol. When employing a 32-shot adaptation protocol, the MAPE error was in the range from 17% to 51%.

The observed high standard deviation in metric values for adaptation set sizes of 2, 4, or 8 samples is attributable to inherent variability in the selection of these sets. This variability indicates that the model's predictions are significantly influenced by the specific compounds included in the adaptation set. To mitigate this impact, a potential strategy involves reducing the number of trainable parameters. It is hypothesized that while structure-toxicity correlations may be shared across different endpoints, their relative significance varies depending on the specific endpoint. Although this assumption simplifies the problem, the potential benefits of parameter reduction may outweigh the drawbacks associated with oversimplification and introduced bias. To validate this hypothesis, a modified variant of MAML, termed Attentive MAML (Att-MAML), was introduced. This framework incorporates an auxiliary scaling layer within the neural network architecture,

designed to modulate the feature importance pertaining to a specific endpoint of interest. Crucially, only the weights associated with this learned scaler are optimized during the inner loop of meta-training. In contrast, all network weights are updated during the outer loop of MAML, enabling the model to acquire general structure-toxicity relationships. During subsequent testing on held-out tasks, the adaptation process adheres to the inner-loop protocol. Consequently, while traditional MAML focuses on learning an adaptation strategy ("learning how to learn"), Att-MAML enables the model to learn how to scale features to adapt the relative importance of structure-toxicity relationships identified during meta-training, effectively adapting the feature importance. A visual representation of this proposed modification is provided in Figure 5.11.

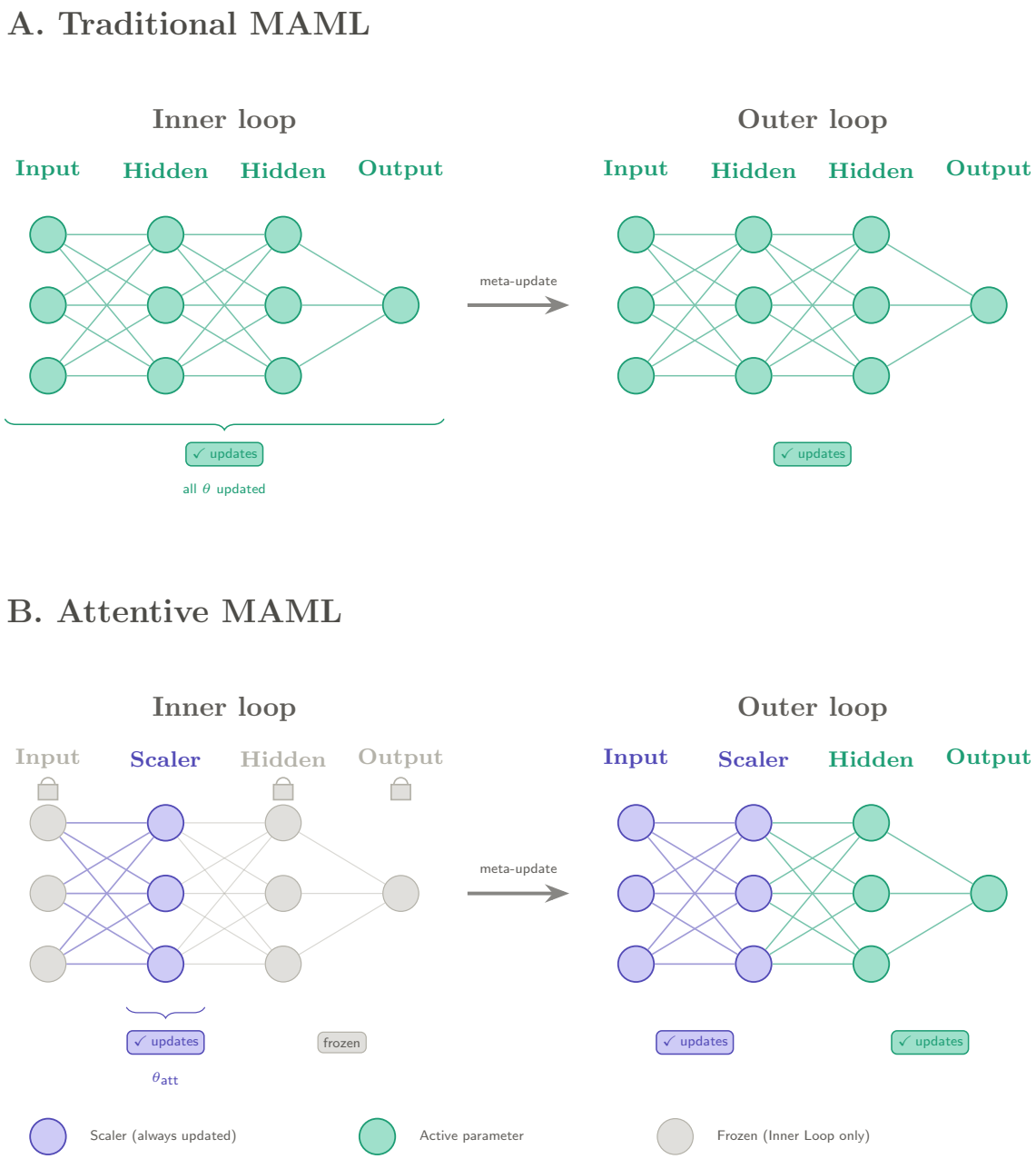


Figure 5.11: Visual explanation of MAML algorithm and the proposed attentive (Att-MAML) modification

The efficacy of the Att-MAML methodology was evaluated across four primary tasks utilizing a leave-one-task-out protocol, in contrast to the conventional MAML approach. As illustrated in Figure 5.12, the comparative analysis between MAML and Att-MAML demonstrates that the modified protocol facilitates a reduction in the standard deviations of performance metrics, thereby diminishing the model's reliance on the selection of adaptation sets. Furthermore, in certain instances, Att-MAML yielded enhanced metric values. However, this improvement was contingent upon an increased computational cost, specifically necessitating a prolongation of the meta-training phase due to the requirement for a greater number of training episodes to effectively extract the structural-toxicity relationships.

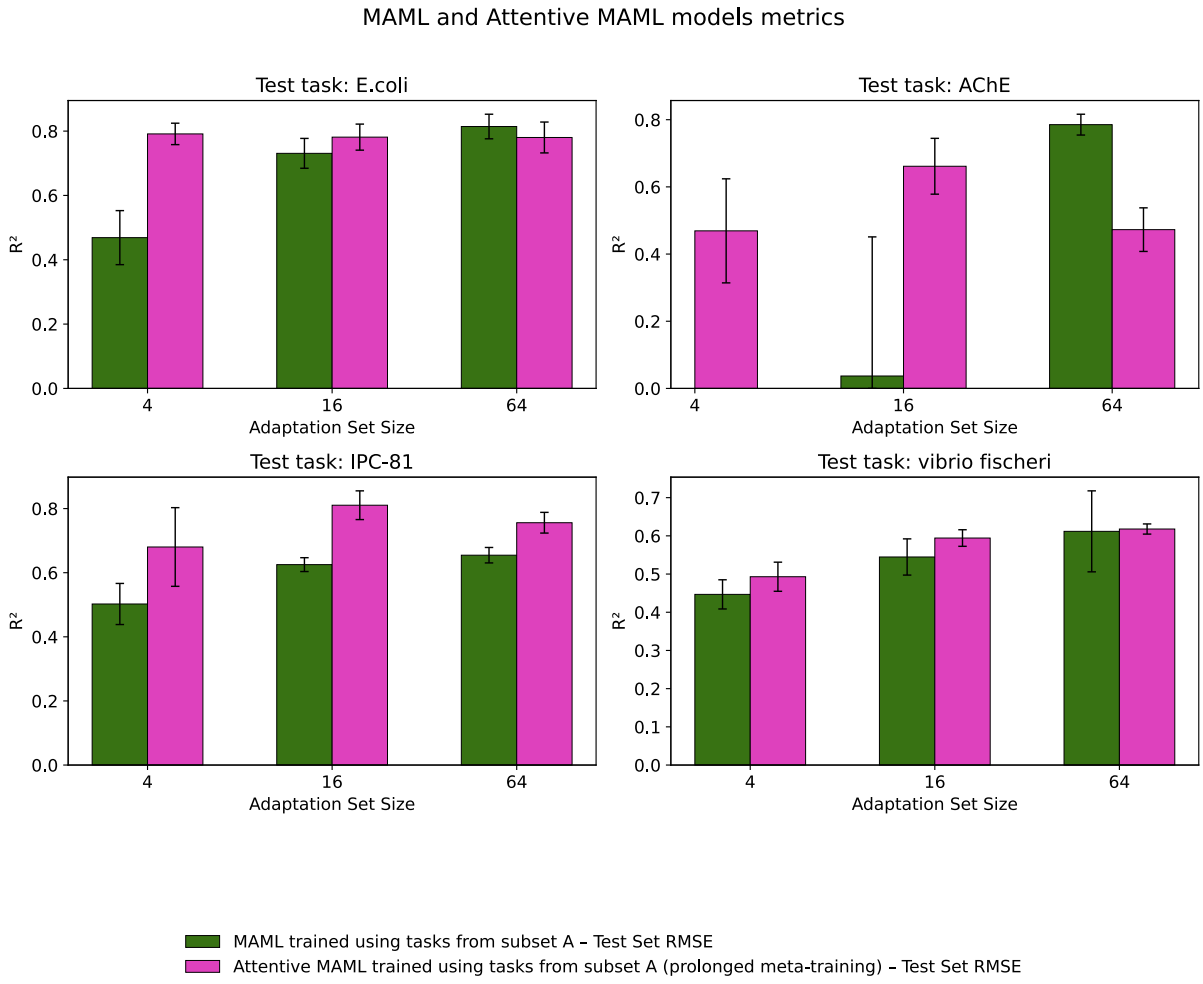


Figure 5.12: Attentive MAML (red) comparison with traditional MAML (green) (negative metrics values not shown)

5.5 Task Similarity Aware Meta-Learning

The application of MAML for modeling the toxicity of ILs yielded valuable insights; however, certain limitations regarding the potential of MAML remain unresolved. Specifically, the absence of an unambiguous task embedding precluded the quantitative evaluation of how task similarity influences MAML performance. A fundamental assumption underpinning this approach is that the MAML meta-model will exhibit superior adaptability

to tasks that are structurally similar to those encountered during meta-training. This assumption is particularly pertinent in the context of solubility prediction, where structurally analogous solutes are expected to exhibit comparable solubility profiles within a given solvent. To empirically test this hypothesis, the Tanimoto similarity between solute molecules was employed as a proxy for task similarity. This definition allowed for the calculation of a pairwise similarity matrix. To assess the degree of similarity of a specific task relative to the broader meta-training set, the maximum similarity value (i.e., the similarity to the closest neighbor in chemical space) was calculated for each task. Consequently, a low value for this metric indicates that no similar compounds exist within the meta-training set, whereas a high value suggests the presence of at least one closely resembling molecule. Based on these derived values, the tasks were subsequently categorized as either similar or dissimilar to the other meta-training tasks.

To investigate the dependency of MAML performance on the similarity between test tasks and meta-training tasks, a series of computational experiments were conducted utilizing tasks from the DS6 dataset. Given that DS6 comprises approximately 110 tasks, a subset of 12 tasks (representing roughly 10% of the total) was selected for testing. To enhance the statistical robustness of the sample, a cross-validation-like protocol was implemented. Specifically, 18 tasks were randomly selected from both the 20 most similar and the 20 most dissimilar tasks. This selection was further structured into three folds for task-cross-validation. In this setup, the MAML model was trained on the meta-training tasks and subsequently adapted using a 128-shot adaptation set for each test task. Each experimental run consisted of 6 similar and 6 dissimilar testing tasks, totaling 18 tasks per group across the three folds. The comparative performance between MAML and a classically trained GNN is presented in Figure 5.13. The relative difference in $RMSE$ error was quantified as $(RMSE_{GNN} - RMSE_{MAML})/RMSE_{GNN}$. The analysis revealed that for the dissimilar task group, the relative error difference occasionally exhibited a larger magnitude. Conversely, for the similar task group, the majority of the relative error differences clustered near zero, suggesting that MAML achieves competitive performance relative to GNN, despite the latter typically utilizing larger sample sizes than 128 per task. Furthermore, a non-parametric Kruskal–Wallis test demonstrated that the difference in the relative $RMSE$ error between the similar and dissimilar task groups was statistically significant ($p = 0.04$).

The suboptimal performance of MAML on tasks dissimilar to the meta-training distribution indicates a significant challenge in out-of-distribution adaptation. This deficiency stems from both the inherent limitations of the MAML algorithm and its lower baseline performance compared to the GNN model. To address this, the Reptile algorithm was adopted, as its simpler meta-learning structure more closely approximates traditional GNN training, potentially yielding superior pre-adaptation performance. While transitioning to Reptile necessitated architectural adjustments, particularly regarding batch normalization stability, the selection was driven by the need for a meta-learning approach that better preserves the structural feature extraction capabilities of the GNN model.

The Reptile algorithm was further modified, as shown in Algorithm 3. This algorithm is designed to facilitate the efficient training of a model, parameterized by θ, across a diverse distribution of tasks, $p(\mathcal{T})$. It constitutes a modification of the standard Reptile algorithm by introducing a "Task Similarity Factor" (TSF) to dynamically modulate the learning rate based on the degree of similarity between the active task and the broader

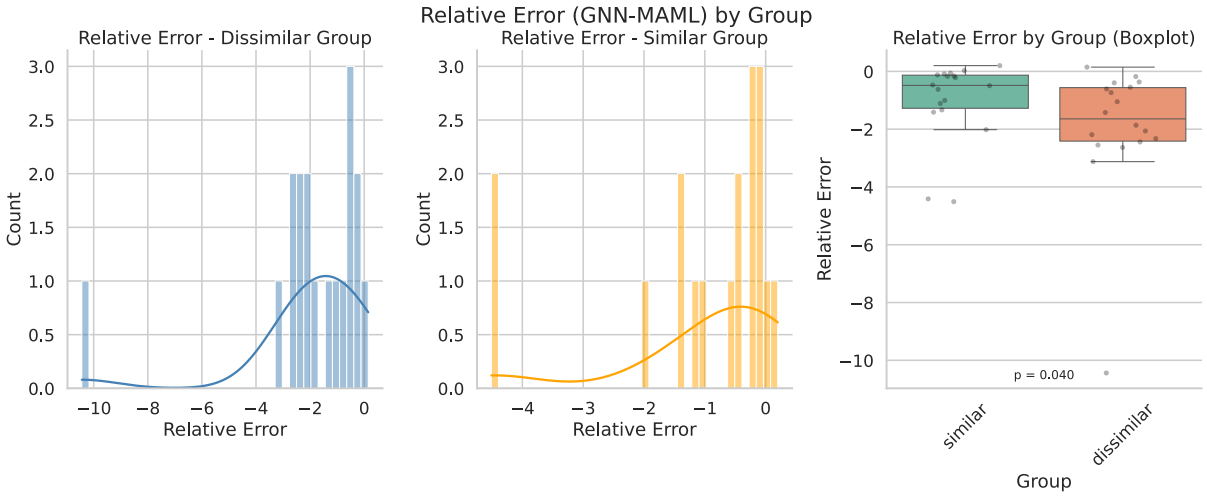


Figure 5.13: Relative error difference between GNN and MAML by group: similar or dissimilar tasks

task distribution. The procedure commences with the stochastic selection of a task, $\mathcal{T}$, from $p(\mathcal{T})$. Subsequently, a data batch, B_k, is sampled exclusively from the selected task $\mathcal{T}$. The similarity between $\mathcal{T}$ and all other tasks within $p(\mathcal{T})$ is then quantified. The maximum similarity score, denoted as τ, is identified, and the TSF is computed via the relation: $TSF = 2 - \max(\tau)$. The final phase involves updating the model parameters θ by performing a gradient descent step on the loss function $\mathcal{L}_{\mathcal{T}}$ associated with the current task and data batch. Critically, this update is scaled by the TSF. These iterative steps continue until the model exhibits convergence, i.e., its performance ceases to improve. The TSF serves as a mechanism to regulate the magnitude of the learning step: when the current task exhibits high similarity to the general distribution (i.e., τ is large), the TSF is small, resulting in a conservative update that mitigates premature over-specialization. Conversely, when the current task is highly unique or dissimilar to the distribution (i.e., τ is small), the TSF is large, enabling a more aggressive adaptation to the specific characteristics of that task.

Algorithm 3 The Task Similarity Aware Reptile algorithm

Require: Task distribution $p(\mathcal{T})$, LR α
Initialize parameters θ
while not converged **do**
  Sample task $\mathcal{T} \sim p(\mathcal{T})$
  for $k = 1$ to K **do**
    Sample data batch B_k from $\mathcal{T}$
    $\tau \leftarrow \{\mathrm{Tc}(\mathcal{T}, \tilde{t}) : \tilde{t} \in p(\mathcal{T})\}$
    $TSF \leftarrow 2 - max(\tau)$
    $\theta \leftarrow \theta - \alpha \nabla_\theta \mathcal{L}_{\mathcal{T}}(\theta; B_k) \cdot TSF$
  end for
end while

The core mechanism of the proposed algorithm involves modulating the loss function by a task similarity factor, as illustrated in Figure 5.14. Conceptually, the procedure entails randomly selecting a task and subsequently computing the similarity of its associated

solute molecule against all other meta-training solutes. The resulting vector of similarity scores is then aggregated into a scalar value via a maximum operation. To enhance computational efficiency, this scalar is pre-computed to circumvent the overhead associated with pairwise calculations during training. This pre-computed value serves as a scaling factor applied to the loss function. Specifically, in its most straightforward linear formulation, the scaling factor is defined as $2 - \max(\text{similarity})$. Consequently, tasks possessing highly similar analogs are assigned a loss value approaching unity, whereas dissimilar tasks are assigned higher weights, thereby promoting a more robust and flexible neural network architecture following the meta-training phase.

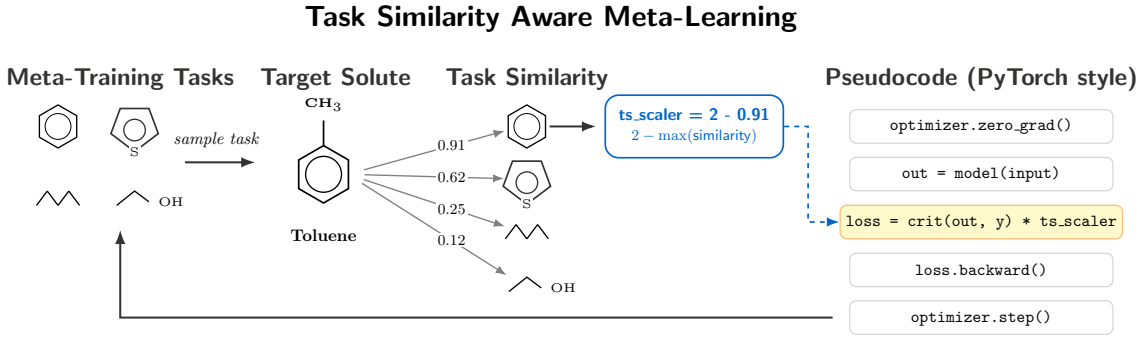


Figure 5.14: Visual explanation of Task Similarity Aware Meta-Learning

The TSA-Reptile model was designed to address the modeling of dissimilar tasks. By employing loss scaling, the network was prevented from overfitting to the adaptation of the average task, thereby facilitating superior adaptation to out-of-distribution tasks relative to the training set. This approach was validated across six distinct tasks, specifically concerning the activity coefficients of 2-propanol, tert-butylethyl ether, diethyl ether, methanol, water, and acetonitrile within the context of ILs (Figure 5.15). The empirical results demonstrated that the TSA-Reptile model exhibited superior performance compared to the MAML algorithm on five of the six evaluated tasks.

The scaling of the loss function in the TSA-Reptile framework necessitates further investigation. Several scaling strategies could be considered. The linear scaling approach, which was previously detailed, represents the most straightforward option. Alternative scaling mechanisms include exponential scaling ($\exp(-S)$) or inverse scaling ($1/S$) with respect to the similarity S to the meta-training task. Figure 5.16 illustrates the comparative performance of these variants against the MAML method across both similar and dissimilar tasks. For tasks exhibiting similarity to the meta-training tasks, the MAML method demonstrates a significant performance advantage over the Reptile-based approaches, suggesting that MAML's capability to adapt to tasks resembling the meta-training set is a contributing factor. Conversely, dissimilar tasks yield divergent results. Across all TSA-Reptile variants employing distinct scaling functions, the performance metrics are largely comparable, with the inverse scaling variant exhibiting a marginal superiority over the other configurations.

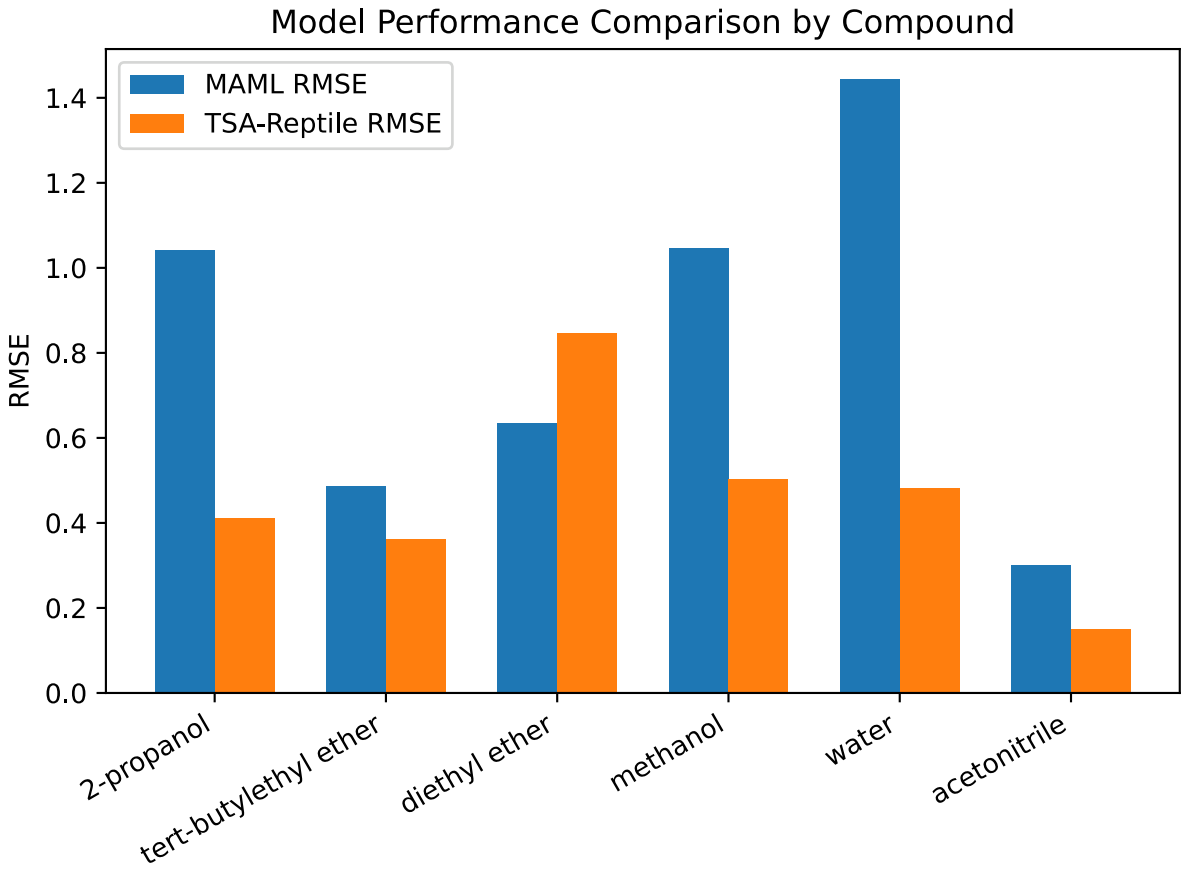


Figure 5.15: Performance Comparison of MAML and TSA-Reptile by Compound

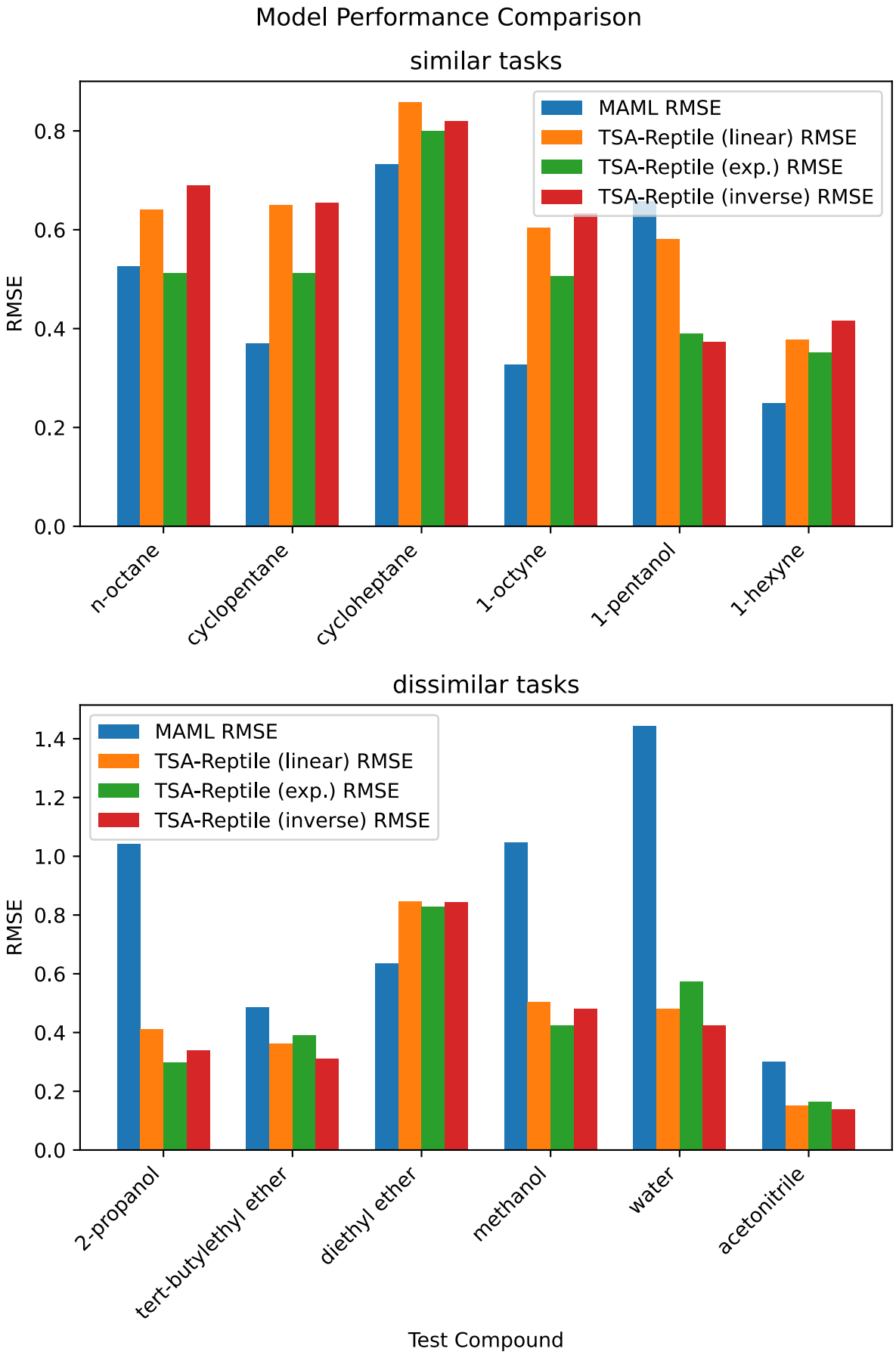


Figure 5.16: Performance Comparison of MAML and TSA-Reptile by Compound

Chapter 6

Conclusions

This dissertation addresses a critical bottleneck in computational chemistry: the optimization of machine learning and deep learning models under data-limited regimes. By addressing the inherent challenges of data scarcity in chemoinformatics, this work delineates distinct pathways for advancing both chemical modeling applications and the underlying computational methodologies. Through a sustained investigation into ionic liquids, deep learning architectures, and meta-learning strategies, it was demonstrated that significant progress can be made even when training data is limited.

The primary application of this research concerns the optimization of scientific research in solvent discovery. It was demonstrated that a novel descriptor could be proposed to significantly enhance the quality of non-linear models for carbon dioxide solubility prediction. Through the utilization of association rules, it was revealed that specific combinations of moieties that favor both microscopic (free volume) and macroscopic (Henry's Law Constant) properties describing sorption. Crucially, this study validates the concept that association rules provide reliable guidance on beneficial ion pairs favoring solubility. Although chemical interpretation was not the sole objective of the primary modeling tasks, it was shown that these show cases offer a validated interpretative tool that can influence future research in areas where model interpretation is an effective means of drawing conclusions. This includes research on lignin isolation, where the interpretability of model predictions served as a validation of the models' rationale. Furthermore, the proposed pipeline proposed is sufficiently robust to serve as a foundational basis for subsequent community efforts on modeling lignin isolation with deep eutectic solvents.

Regarding the computational methods, a major effort was devoted to deep learning approaches for bimolecular systems. While graph priors remain influential in chemical modeling, a gap in how to apply them to bimolecular systems was addressed. This work involved reviewing the literature and benchmarking known and newly proposed methods to represent ionic liquids as graphs, alongside an evaluation of data quality.

Regarding the meta-learning approaches, domain-specific adaptations to the meta-learning approaches have been proposed. In toxicity prediction, introduction of a bias to MAML based on structure-toxicity shared between endpoints was introduced. This resulted in a model that learns only latent feature importance during adaptation. Similarly, in sorption modeling, it was suggested that adapting a model is easier when similar tasks (as approximated via the Tanimoto coefficient between solutes) are utilized in meta-training.

Consequently, task similarity-based scaling of the loss function was implemented to increase the performance of the Reptile model on dissimilar, out-of-distribution tasks.

Ultimately, this research contributes to both the applications and methods of chemoinformatics. I demonstrate that leveraging chemical similarity between target molecules serves as a powerful proxy for task similarity, resolving the challenge of assessing similarity in tasks that vary significantly. By improving meta-learning approaches, it was demonstrated how the structure of bimolecular systems should be expressed with graphs. These contributions validate that low-data ML is not only feasible for complex molecular systems but also capable of providing explainable insights that align with chemical intuition. This thesis proves that with the right architectural and learning adjustments, deep learning can offer robust, interpretable tools for solving emerging challenges in chemistry.

List of publications forming the series

A1. **Baran, K.***, & Kloskowski, A. (2023). Graph Neural Networks and Structural Information on Ionic Liquids: A Cheminformatics Study on Molecular Physicochemical Property Prediction. Journal of Physical Chemistry B, 127, 10542-10555. doi.org/10.1021/acs.jpcb.3c05521

A2. **Baran, K.***, Barczak, B., & Kloskowski, A. (2024). Modeling lignin extraction with ionic liquids using machine learning approach. Science of the Total Environment, 935, 173234-173248. doi.org/10.1016/j.scitotenv.2024.173234

A3. **Baran, K.***, & Kloskowski, A. (2025). Unlocking Structural Insights into CO2 Absorption with Ionic Liquids: A Comparative Study of Machine Learning, Association Rules, and Meta-Learning for Modeling Henry's Law Constant. ACS Sustainable Chemistry & Engineering, 13, 12805-12817. doi.org/10.1021/acssuschemeng.5c06122

A4. **Baran, K.***, Derron, T., Eichenlaub, J., & Kloskowski, A. (2026). Few-Shot Prediction of Toxicity of Ionic Liquids Supported by Attentive Model-Agnostic Meta-Learning. Chemical Research in Toxicology. doi.org/10.1021/acs.chemrestox.6c00022

A5. **Baran, K.***, & Kloskowski, A. (2026). Exploring Chemoinformatics Aspects of Few-Shot Meta-Learning by Example of an Infinite Dilution Activity Coefficient in Ionic Liquid Prediction. Journal of Chemical Information and Modeling. doi.org/10.1021/acs.jcim.6c00067

* corresponding author

Acknowledgments

The author would like to gratefully acknowledge that this research was funded by the National Science Centre, Poland (NCN) under the Preludium 22 program in the years 2024–2027 (project no. UMO-2023/49/N/ST5/01043, principal investigator:**Karol Baran**). Poland's high-performance Infrastructure, PLGrid HPC Centers: ACK Cyfronet AGH, PCSS, CI TASK, WCSS (Grant No. PLG/2024/017245, principal investigator: **Karol Baran**), and the Centre of Informatics Tricity Academic Supercomputer & Network (Grant No. PT00996, principal investigator: **Karol Baran**), are gratefully acknowledged for providing the computational resources used for carrying out calculations in this study.

List of Figures

List of Tables

List of Algorithms

List of Equations

Bibliography

[1] N. D. Lawrence and J. Montgomery, "Accelerating ai for science: Open data science for science," *Royal Society Open Science*, vol. 11, no. 8, p. 231 130, 2024.

[2] L. Fanti, D. Guarascio, and M. Moggi, "From heron of alexandria to amazon's alexa: A stylized history of ai and its impact on business models, organization and work," *Journal of Industrial and Business Economics*, vol. 49, no. 3, pp. 409–440, 2022.

[3] J. Esparza and M. Blondin, *Automata theory: An algorithmic approach*. MIT Press, 2023.

[4] B. P. Bhuyan, A. Ramdane-Cherif, T. P. Singh, and R. Tomar, "Rule-based systems and expert systems," in *Neuro-Symbolic Artificial Intelligence: Bridging Logic and Learning*, Springer, 2024, pp. 63–85.

[5] Y. Bengio, Y. Lecun, and G. Hinton, "Deep learning for ai," *Communications of the ACM*, vol. 64, no. 7, pp. 58–65, 2021.

[6] J. J. Hopfield, "Neural networks and physical systems with emergent collective computational abilities.," *Proceedings of the National Academy of Sciences*, vol. 79, no. 8, pp. 2554–2558, Apr. 1982, ISSN: 1091-6490. DOI: 10.1073/pnas.79.8.2554.

[7] H. Alaskar and T. Saba, "Machine learning and deep learning: A comparative review," *Proceedings of Integrated Intelligence Enable Networks and Computing: IIENC 2020*, pp. 143–150, 2021.

[8] Y. Roh, G. Heo, and S. E. Whang, "A survey on data collection for machine learning: A big data-ai integration perspective," *IEEE Transactions on Knowledge and Data Engineering*, vol. 33, no. 4, pp. 1328–1347, 2019.

[9] B. Khosravi, A. D. Weston, F. Nugen, *et al.*, "Demystifying statistics and machine learning in analysis of structured tabular data," *The Journal of arthroplasty*, vol. 38, no. 10, pp. 1943–1947, 2023.

[10] M. Alloghani, D. Al-Jumeily, J. Mustafina, A. Hussain, and A. J. Aljaaf, "A systematic review on supervised and unsupervised machine learning algorithms for data science," *Supervised and unsupervised learning for data science*, pp. 3–21, 2020.

[11] C. Grosan and A. Abraham, "Hybrid intelligent systems," in *Intelligent Systems: A Modern Approach*, Springer, 2011, pp. 423–450.

[12] D. Foead, A. Ghifari, M. B. Kusuma, N. Hanafiah, and E. Gunawan, "A systematic literature review of a* pathfinding," *Procedia Computer Science*, vol. 179, pp. 507–514, 2021, ISSN: 1877-0509. DOI: 10.1016/j.procs.2021.01.034.

[13] S. Katoch, S. S. Chauhan, and V. Kumar, "A review on genetic algorithm: Past, present, and future," *Multimedia Tools and Applications*, vol. 80, no. 5, pp. 8091–8126, Oct. 2020, ISSN: 1573-7721. DOI: 10.1007/s11042-020-10139-6.

[14] K. Taunk, S. De, S. Verma, and A. Swetapadma, "A brief review of nearest neighbor algorithm for learning and classification," in *2019 International Conference on Intelligent Computing and Control Systems (ICCS)*, IEEE, May 2019, pp. 1255–1260. DOI: 10.1109/iccs45141.2019.9065747.

[15] O. I. Abiodun, A. Jantan, A. E. Omolara, K. V. Dada, N. A. Mohamed, and H. Arshad, "State-of-the-art in artificial neural network applications: A survey," *Heliyon*, vol. 4, no. 11, e00938, Nov. 2018, ISSN: 2405-8440. DOI: 10.1016/j.heliyon.2018.e00938.

[16] F. Scarselli and A. Chung Tsoi, "Universal approximation using feedforward neural networks: A survey of some existing methods, and some new results," *Neural Networks*, vol. 11, no. 1, pp. 15–37, Jan. 1998, ISSN: 0893-6080. DOI: 10.1016/s0893-6080(97)00097-x.

[17] A. Mahyar, H. Motamednia, P. Cheraaqee, and A. Mansouri, "Feature extraction and deep learning," in *Dimensionality Reduction in Machine Learning*. Elsevier, 2025, pp. 211–243, ISBN: 9780443328183. DOI: 10.1016/b978-0-44-332818-3.00019-8.

[18] Q. Zhang, L. T. Yang, Z. Chen, and P. Li, "A survey on deep learning for big data," *Information Fusion*, vol. 42, pp. 146–157, Jul. 2018, ISSN: 1566-2535. DOI: 10.1016/j.inffus.2017.10.006.

[19] G. Vrbančič and V. Podgorelec, "Transfer learning with adaptive fine-tuning," *Ieee Access*, vol. 8, pp. 196 197–196 211, 2020.

[20] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *Nature*, vol. 521, no. 7553, pp. 436–444, May 2015, ISSN: 1476-4687. DOI: 10.1038/nature14539.

[21] P. Ramachandran, N. Parmar, A. Vaswani, I. Bello, A. Levskaya, and J. Shlens, "Stand-alone self-attention in vision models," in *Advances in Neural Information Processing Systems*, H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, Eds., vol. 32, Curran Associates, Inc., 2019. [Online]. Available: https://proceedings.neurips.cc/paper_files/paper/2019/file/3416a75f4cea9109507cacd8e2f2aefc-Paper.pdf.

[22] D. E. Rumelhart, G. E. Hinton, and R. J. Williams, "Learning internal representations by error propagation," Tech. Rep., 1985.

[23] F. Scarselli, M. Gori, A. C. Tsoi, M. Hagenbuchner, and G. Monfardini, "The graph neural network model," *IEEE Transactions on Neural Networks*, vol. 20, no. 1, pp. 61–80, Jan. 2009, ISSN: 1941-0093. DOI: 10.1109/tnn.2008.2005605.

[24] J. R. Quinlan, "Induction of decision trees," *Machine Learning*, vol. 1, no. 1, pp. 81–106, Mar. 1986, ISSN: 1573-0565. DOI: 10.1007/bf00116251.

[25] L. Breiman, "Random forests," *Machine Learning*, vol. 45, no. 1, pp. 5–32, Oct. 2001, ISSN: 1573-0565. DOI: 10.1023/a:1010933404324.

[26] C. Bentéjac, A. Csörgő, and G. Martınez-Muñoz, "A comparative analysis of gradient boosting algorithms," *Artificial Intelligence Review*, vol. 54, no. 3, pp. 1937–1967, 2021.

[27] S. Bhat and P. Szczuko, "Impact of canny edge detection preprocessing on performance of machine learning models for parkinson's disease classification," *Scientific Reports*, vol. 15, no. 1, p. 16 413, 2025.

[28] T. P. Lillicrap, A. Santoro, L. Marris, C. J. Akerman, and G. Hinton, "Backpropagation and the brain," *Nature Reviews Neuroscience*, vol. 21, no. 6, pp. 335–346, 2020.

[29] A. N. Elmachtoub, J. C. N. Liang, and R. McNellis, "Decision trees for decision-making under the predict-then-optimize framework," in *Proceedings of the 37th International Conference on Machine Learning*, ser. ICML'20, JMLR.org, 2020.

[30] D. T. Larose and C. D. Larose, *Discovering Knowledge in Data: An Introduction to Data Mining*. Wiley, Jun. 2014, ISBN: 9781118874059. DOI: 10.1002/9781118874059.

[31] K. Tyagi, C. Rane, R. Sriram, and M. Manry, "Unsupervised learning," in *Artificial Intelligence and Machine Learning for EDGE Computing*. Elsevier, 2022, pp. 33–52, ISBN: 9780128240540. DOI: 10.1016/b978-0-12-824054-0.00012-5.

[32] C. Miller, T. Portlock, D. M. Nyaga, and J. M. O'Sullivan, "A review of model evaluation metrics for machine learning in genetics and genomics," *Frontiers in bioinformatics*, vol. 4, p. 1 457 619, 2024.

[33] L.-b. Sweet, C. Müller, M. Anand, and J. Zscheischler, "Cross-validation strategy impacts the performance and interpretation of machine learning models," *Artificial Intelligence for the Earth Systems*, vol. 2, no. 4, e230026, 2023.

[34] T. Burzykowski, M. Geubbelmans, A.-J. Rousseau, and D. Valkenborg, "Validation of machine learning algorithms," *American journal of orthodontics and dentofacial orthopedics*, vol. 164, no. 2, pp. 295–297, 2023.

[35] N. Głowacka and J. Rumiński, "Face with mask detection in thermal images using deep neural networks," *Sensors*, vol. 21, no. 19, p. 6387, 2021.

[36] H. Xu, Q. Xu, F. Cong, *et al.*, "Vision transformers for computational histopathology," *IEEE Reviews in Biomedical Engineering*, vol. 17, pp. 63–79, 2023.

[37] OpenAI, J. Achiam, S. Adler, *et al.*, *Gpt-4 technical report*, 2024. arXiv: 2303.08774 [cs.CL]. [Online]. Available: https://arxiv.org/abs/2303.08774.

[38] W. Fan, Y. Ma, Q. Li, *et al.*, "Graph neural networks for social recommendation," in *The World Wide Web Conference*, ser. WWW '19, ACM, May 2019, pp. 417–426. DOI: 10.1145/3308558.3313488.

[39] M. van Aalst, T. Nies, T. Pfennig, and A. Matuszyńska, "Mxlpy—python package for mechanistic learning and hybrid modelling in life science," *Bioinformatics Advances*, vol. 5, no. 1, vbaf294, 2025.

[40] J. Jumper, R. Evans, A. Pritzel, *et al.*, "Highly accurate protein structure prediction with alphafold," *nature*, vol. 596, no. 7873, pp. 583–589, 2021.

[41] A. Varnek and I. I. Baskin, "Chemoinformatics as a theoretical chemistry discipline," *Molecular Informatics*, vol. 30, no. 1, pp. 20–32, 2011.

[42] D. S. Wigh, J. M. Goodman, and A. A. Lapkin, "A review of molecular representation in the age of machine learning," *Wiley Interdisciplinary Reviews: Computational Molecular Science*, vol. 12, no. 5, e1603, 2022.

[43] M. E. Mswahili and Y.-S. Jeong, "Transformer-based models for chemical smiles representation: A comprehensive literature review," *Heliyon*, vol. 10, no. 20, 2024.

[44] M. Krenn, Q. Ai, S. Barthel, *et al.*, "Selfies and the future of molecular string representations," *Patterns*, vol. 3, no. 10, 2022.

[45] J. Shen and C. A. Nicolaou, "Molecular property prediction: Recent trends in the era of artificial intelligence," *Drug Discovery Today: Technologies*, vol. 32, pp. 29–36, 2019.

[46] J. Deng, Z. Yang, H. Wang, I. Ojima, D. Samaras, and F. Wang, "A systematic study of key elements underlying molecular property prediction," *Nature Communications*, vol. 14, no. 1, p. 6395, 2023.

[47] E. N. Muratov, J. Bajorath, R. P. Sheridan, *et al.*, "Qsar without borders," *Chemical Society Reviews*, vol. 49, no. 11, pp. 3525–3564, 2020.

[48] A. Storkey *et al.*, "When training and test sets are different: Characterizing learning transfer," *Dataset shift in machine learning*, vol. 30, no. 3-28, p. 6, 2009.

[49] J. Deng, Z. Yang, H. Wang, I. Ojima, D. Samaras, and F. Wang, "A systematic study of key elements underlying molecular property prediction," *Nature Communications*, vol. 14, no. 1, Oct. 2023, ISSN: 2041-1723. DOI: 10.1038/s41467-023-41948-6.

[50] J.-L. Reymond, "The chemical space project," *Accounts of chemical research*, vol. 48, no. 3, pp. 722–730, 2015.

[51] A. Mauri, V. Consonni, R. Todeschini, *et al.*, "Molecular descriptors," in *Handbook of computational chemistry*, Springer International Publishing, 2017, pp. 2065–2093.

[52] D. Rogers and M. Hahn, "Extended-connectivity fingerprints," *Journal of Chemical Information and Modeling*, vol. 50, no. 5, pp. 742–754, Apr. 2010, ISSN: 1549-960X. DOI: 10.1021/ci100050t.

[53] S. Zhong and X. Guan, "Count-based morgan fingerprint: A more efficient and interpretable molecular representation in developing machine learning-based predictive regression models for water contaminants' activities and properties," *Environmental science & technology*, vol. 57, no. 46, pp. 18 193–18 202, 2023.

[54] D. K. Duvenaud, D. Maclaurin, J. Iparraguirre, *et al.*, "Convolutional networks on graphs for learning molecular fingerprints," *Advances in neural information processing systems*, vol. 28, 2015.

[55] N. C. Chung, B. Miasojedow, M. Startek, and A. Gambin, "Jaccard/tanimoto similarity test and estimation methods for biological presence-absence data," *BMC bioinformatics*, vol. 20, no. Suppl 15, p. 644, 2019.

[56] P. Willett, "Similarity-based virtual screening using 2d fingerprints," *Drug discovery today*, vol. 11, no. 23-24, pp. 1046–1053, 2006.

[57] B. Selvaratnam and R. T. Koodali, "Machine learning in experimental materials chemistry," *Catalysis Today*, vol. 371, pp. 77–84, 2021.

[58] W. P. Walters, M. T. Stahl, and M. A. Murcko, "Virtual screening—an overview," *Drug discovery today*, vol. 3, no. 4, pp. 160–178, 1998.

[59] V. Bagal, R. Aggarwal, P. Vinod, and U. D. Priyakumar, "Molgpt: Molecular generation using a transformer-decoder model," *Journal of chemical information and modeling*, vol. 62, no. 9, pp. 2064–2076, 2021.

[60] A. Racki and K. Paduszynski, "Recent advances in the modeling of ionic liquids using artificial neural networks," *Journal of Chemical Information and Modeling*, vol. 65, no. 7, pp. 3161–3175, 2025.

[61] Z. Lei, B. Chen, Y.-M. Koo, and D. R. MacFarlane, *Introduction: Ionic liquids*, 2017.

[62] K. N. Marsh, J. A. Boxall, and R. Lichtenthaler, "Room temperature ionic liquids and their mixtures—a review," *Fluid phase equilibria*, vol. 219, no. 1, pp. 93–98, 2004.

[63] N. V. Plechkova and K. R. Seddon, "Ionic liquids:"designer" solvents for green chemistry," *Methods and reagents for green chemistry: an introduction*, pp. 103–130, 2007.

[64] Z. Song, J. Chen, J. Cheng, G. Chen, and Z. Qi, "Computer-aided molecular design of ionic liquids as advanced process media: A review from fundamentals to applications," *Chemical Reviews*, vol. 124, no. 2, pp. 248–317, 2023.

[65] A. Rybińska-Fryca, A. Sosnowska, and T. Puzyn, "Representation of the structure—a key point of building qsar/qspr models for ionic liquids," *Materials*, vol. 13, no. 11, p. 2500, 2020.

[66] B. Sepehri, "A review on created qspr models for predicting ionic liquids properties and their reliability from chemometric point of view," *Journal of Molecular Liquids*, vol. 297, p. 112 013, 2020.

[67] D. D. Patel and J.-M. Lee, "Applications of ionic liquids," *The Chemical Record*, vol. 12, no. 3, pp. 329–355, 2012.

[68] K. Paduszyński, "Extensive databases and group contribution qsprs of ionic liquids properties. 1. density," *Industrial & Engineering Chemistry Research*, vol. 58, no. 13, pp. 5322–5338, Mar. 2019, ISSN: 1520-5045. DOI: 10.1021/acs.iecr.9b00130.

[69] Y. Jian, Y. Wang, and A. Barati Farimani, "Predicting co2 absorption in ionic liquids with molecular descriptors and explainable graph neural networks," *ACS Sustainable Chemistry & Engineering*, vol. 10, no. 50, pp. 16 681–16 691, 2022.

[70] W. Beckner, C. Ashraf, J. Lee, D. A. Beck, and J. Pfaendtner, "Continuous molecular representations of ionic liquids," *The Journal of Physical Chemistry B*, vol. 124, no. 38, pp. 8347–8357, 2020.

[71] G. Chen, Z. Song, Z. Qi, and K. Sundmacher, "Generalizing property prediction of ionic liquids from limited labeled data: A one-stop framework empowered by transfer learning," *Digital Discovery*, vol. 2, no. 3, pp. 591–601, 2023.

[72] W. Silva, M. Zanatta, A. S. Ferreira, M. C. Corvo, and E. J. Cabrita, "Revisiting ionic liquid structure-property relationship: A critical analysis," *International journal of molecular sciences*, vol. 21, no. 20, p. 7745, 2020.

[73] X. Chen, G. Chen, K. Xie, *et al.*, "Exploring the chemical space of ionic liquids for co2 dissolution through generative machine learning models," *Green Chemical Engineering*, vol. 6, no. 3, pp. 335–343, Sep. 2025, ISSN: 2666-9528. DOI: 10.1016/j.gce.2024.06.005.

[74] D. Horvath, G. Marcou, and A. Varnek, "Generative topographic mapping approach to chemical space analysis," in *Advances in QSAR Modeling: Applications in Pharmaceutical, Chemical, Food, Agricultural and Environmental Sciences*, Springer, 2017, pp. 167–199.

[75] A. Steffen, T. Kogej, C. Tyrchan, and O. Engkvist, "Comparison of molecular fingerprint methods on the basis of biological profile data," *Journal of chemical information and modeling*, vol. 49, no. 2, pp. 338–347, 2009.

[76] A. Varnek, D. Fourches, D. Horvath, *et al.*, "Isida-platform for virtual screening based on fragment and pharmacophoric descriptors," *Current Computer-Aided Drug Design*, vol. 4, no. 3, p. 191, 2008.

[77] N. Tsuji, P. Sidorov, C. Zhu, *et al.*, "Predicting highly enantioselective catalysts using tunable fragment descriptors," *Angewandte Chemie*, vol. 135, no. 11, e202218659, 2023.

[78] S. Zhou, J. Li, Z. Liu, and D. Wang, "Molecular representation matters: Comparative evaluation of fingerprints, rdkit descriptors, and hashing effects," *The Journal of Physical Chemistry Letters*, 2026.

[79] A. Mauri, V. Consonni, M. Pavan, R. Todeschini, *et al.*, "Dragon software: An easy approach to molecular descriptor calculations," *Match*, vol. 56, no. 2, pp. 237–248, 2006.

[80] A. Mauri and M. Bertola, "Alvascience: A new software suite for the qsar workflow applied to the blood–brain barrier permeability," *International Journal of Molecular Sciences*, vol. 23, no. 21, p. 12 882, 2022.

[81] M. Randić, B. Jerman-Blazić, and N. Trinajstić, "Development of 3-dimensional molecular descriptors," *Computers & chemistry*, vol. 14, no. 3, pp. 237–246, 1990.

[82] F. Berenger, A. Voet, X. Y. Lee, and K. Y. Zhang, "A rotation-translation invariant molecular descriptor of partial charges and its use in ligand-based virtual screening," *Journal of Cheminformatics*, vol. 6, no. 1, p. 23, 2014.

[83] S. Alagarsamy, C.-S. Shieh, M.-F. Horng, S. Radhakrishnan, S. Radhakrishnan, and A. Omri, "Transformers in drug discovery: Fine-tuning chemberta for high-accuracy prediction of solubility, toxicity and binding affinity," *Drug Discovery Today*, p. 104 602, 2026.

[84] S. Koutsoukos, F. Philippi, F. Malaret, and T. Welton, "A review on machine learning algorithms for the ionic liquid chemical space," *Chemical science*, vol. 12, no. 20, pp. 6820–6843, 2021.

[85] J. Eichenlaub, P. W. Rakowska, and A. Kloskowski, "User-assisted methodology targeted for building structure interpretable qspr models for boosting co2 capture with ionic liquids," *Journal of Molecular Liquids*, vol. 350, p. 118 511, 2022.

[86] M. Amiri and F. Khaleseh, "Predicting drug solubility in binary solvent mixtures using graph convolutional networks: A comprehensive deep learning approach," *Scientific Reports*, 2025.

[87] Z. Wu, B. Ramsundar, E. N. Feinberg, *et al.*, "Moleculenet: A benchmark for molecular machine learning," *Chemical Science*, vol. 9, no. 2, pp. 513–530, 2018, ISSN: 2041-6539. DOI: 10.1039/c7sc02664a.

[88] M. Stanley, J. F. Bronskill, K. Maziarz, *et al.*, "Fs-mol: A few-shot learning dataset of molecules," in *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 2)*, 2021.

[89] K. Huang, T. Fu, W. Gao, *et al.*, "Therapeutics data commons: Machine learning datasets and tasks for drug discovery and development," *arXiv preprint arXiv:2102.09548*, 2021.

[90] K. O. Klimenko, J. M. Inês, J. M. S. S. Esperança, L. P. N. Rebelo, J. Aires-de-Sousa, and G. V. S. M. Carrera, "Qspr modeling of liquid-liquid equilibria in two-phase systems of water and ionic liquid," *Molecular informatics*, vol. 39, no. 9, p. 2 000 001, 2020.

[91] Q. Dong, C. D. Muzny, A. Kazakov, *et al.*, "Ilthermo: A free-access web database for thermodynamic properties of ionic liquids," *Journal of Chemical & Engineering Data*, vol. 52, no. 4, pp. 1151–1159, 2007.

[92] V. Svetnik, T. Wang, C. Tong, A. Liaw, R. P. Sheridan, and Q. Song, "Boosting: An ensemble learning tool for compound classification and qsar modeling," *Journal of Chemical Information and Modeling*, vol. 45, no. 3, pp. 786–799, Apr. 2005, ISSN: 1549-960X. DOI: 10.1021/ci0500379.

[93] J. Shi, W. Zheng, X. Yuan, Y. Wei, and K. Zhang, "Quantitative structure activity relationship–based prediction of acute exposure guideline levels for aliphatic compounds," *Environmental Toxicology and Chemistry*, vol. 44, no. 8, pp. 2310–2321, May 2025, ISSN: 1552-8618. DOI: 10.1093/etojnl/vgaf137.

[94] A. Wiriyarattanakul, W. Xie, B. Toopradab, *et al.*, "Comparative study of machine learning-based qsar modeling of anti-inflammatory compounds from durian extraction," *ACS Omega*, Feb. 2024, ISSN: 2470-1343. DOI: 10.1021/acsomega.3c07386.

[95] P. Ambure, A. Gajewicz-Skretna, M. N. D. Cordeiro, and K. Roy, "New workflow for qsar model development from small data sets: Small dataset curator and small dataset modeler. integration of data curation, exhaustive double cross-validation, and a set of optimal model selection techniques," *Journal of chemical information and modeling*, vol. 59, no. 10, pp. 4070–4076, 2019.

[96] G. J. Lavado, D. Baderna, E. Carnesecchi, *et al.*, "Qsar models for soil ecotoxicity: Development and validation of models to predict reproductive toxicity of organic chemicals in the collembola folsomia candida," *Journal of Hazardous Materials*, vol. 423, p. 127 236, 2022.

[97] S. Zhang, A. Golbraikh, S. Oloff, H. Kohn, and A. Tropsha, "A novel automated lazy learning qsar (all-qsar) approach: Method development, applications, and virtual screening of chemical databases using validated all-qsar models," *Journal of Chemical Information and Modeling*, vol. 46, no. 5, pp. 1984–1995, Sep. 2006, ISSN: 1549-960X. DOI: 10.1021/ci060132x.

[98] I. Billard, G. Marcou, A. Ouadi, and A. Varnek, "In silico design of new ionic liquids based on quantitative structure-property relationship models of ionic liquid viscosity," *The Journal of Physical Chemistry B*, vol. 115, no. 1, pp. 93–98, Dec. 2010, ISSN: 1520-5207. DOI: 10.1021/jp107868w.

[99] A. Varnek, N. Kireeva, I. V. Tetko, I. I. Baskin, and V. P. Solov'ev, "Exhaustive qspr studies of a large diverse set of ionic liquids: How accurately can we predict melting points?" *Journal of Chemical Information and Modeling*, vol. 47, no. 3, pp. 1111–1122, Mar. 2007, ISSN: 1549-960X. DOI: 10.1021/ci600493x.

[100] V. Venkatraman, S. Evjen, H. K. Knuutila, A. Fiksdahl, and B. K. Alsberg, "Predicting ionic liquid melting points using machine learning," *Journal of Molecular Liquids*, vol. 264, pp. 318–326, Aug. 2018, ISSN: 0167-7322. DOI: 10.1016/j.molliq.2018.03.090.

[101] I. Euldji, W. Benmouloud, K. Paduszynski, C. Si-Moussa, and O. Benkortbi, "Hybrid improved grey wolf support vector regression algorithm for modeling solubilities of apis in pure ionic liquids: σ-profile descriptors," *Journal of Chemical Information and Modeling*, vol. 64, no. 4, pp. 1361–1376, 2024.

[102] X. Kang, Y. Zhao, and Z. Chen, "Atom surface fragment contribution method for predicting the toxicity of ionic liquids," *Journal of Hazardous Materials*, vol. 421, p. 126 705, Jan. 2022, ISSN: 0304-3894. DOI: 10.1016/j.jhazmat.2021.126705.

[103] H. Abdellatif, M. Laidi, C. Si-moussa, A. Amrane, I. Euldji, and W. Benmouloud, "Contributions to the development of prediction models for the toxicity of ionic liquids," *Structural Chemistry*, vol. 36, no. 3, pp. 865–886, Nov. 2024, ISSN: 1572-9001. DOI: 10.1007/s11224-024-02411-4.

[104] I. Semenyuta, V. Kovalishyn, D. Hodyna, Y. Startseva, S. Rogalsky, and L. Metelytsia, "New qstr models to evaluation of imidazolium- and pyridinium-contained ionic liquids toxicity," *Computational Toxicology*, vol. 30, p. 100 309, Jun. 2024, ISSN: 2468-1113. DOI: 10.1016/j.comtox.2024.100309.

[105] J. Yan, G. Liu, H. Chen, *et al.*, "Iltox: A curated toxicity database for machine learning and design of environmentally friendly ionic liquids," *Environmental Science & Technology Letters*, vol. 10, no. 11, pp. 983–988, 2023.

[106] R. Shan, R. Zhang, Y. Gao, *et al.*, "Evaluating ionic liquid toxicity with machine learning and structural similarity methods," *Green Chemical Engineering*, vol. 6, no. 2, pp. 249–262, 2025.

[107] D. Fan, K. Xue, R. Zhang, *et al.*, "Application of interpretable machine learning models to improve the prediction performance of ionic liquids toxicity," *Science of The Total Environment*, vol. 908, p. 168 168, 2024.

[108] Z. Wang, Z. Song, and T. Zhou, "Machine learning for ionic liquid toxicity prediction," *Processes*, vol. 9, no. 1, p. 65, 2020.

[109] F. Haijun, L. Jiajia, and Z. Jian, "Prediction of ionic liquid toxicity by interpretable machine learning," *Chinese Journal of Chemical Engineering*, 2025.

[110] E. Heid, K. P. Greenman, Y. Chung, *et al.*, "Chemprop: A machine learning package for chemical property prediction," *Journal of Chemical Information and Modeling*, vol. 64, no. 1, pp. 9–17, Dec. 2023, ISSN: 1549-960X. DOI: 10.1021/acs.jcim.3c01250.

[111] J. W. Burns, A. S. Zalte, C. R. A. Abreu, *et al.*, *Deep learning foundation models from classical molecular descriptors*, 2026. arXiv: 2506.15792 [cs.LG]. [Online]. Available: https://arxiv.org/abs/2506.15792.

[112] R. Gómez-Bombarelli, J. N. Wei, D. Duvenaud, *et al.*, "Automatic chemical design using a data-driven continuous representation of molecules," *ACS central science*, vol. 4, no. 2, pp. 268–276, 2018.

[113] W. Jin, R. Barzilay, and T. Jaakkola, "Junction tree variational autoencoder for molecular graph generation," in *International conference on machine learning*, PMLR, 2018, pp. 2323–2332.

[114] M. J. Kusner, B. Paige, and J. M. Hernández-Lobato, "Grammar variational autoencoder," in *International conference on machine learning*, PMLR, 2017, pp. 1945–1954.

[115] S. Sadaghiyanfam, H. Kamberaj, and Y. Isler, "Leveraging chemberta and machine learning for accurate toxicity prediction of ionic liquids," *Journal of the Taiwan Institute of Chemical Engineers*, vol. 171, p. 106 030, Jun. 2025, ISSN: 1876-1070. DOI: 10.1016/j.jtice.2025.106030.

[116] J. G. Rittig, K. B. Hicham, A. M. Schweidtmann, M. Dahmen, and A. Mitsos, "Graph neural networks for temperature-dependent activity coefficient prediction of solutes in ionic liquids," *Computers & Chemical Engineering*, vol. 171, p. 108 153, 2023.

[117] D. Fan, K. Xue, Y. Liu, *et al.*, "Modeling the toxicity of ionic liquids based on deep learning method," *Computers & Chemical Engineering*, vol. 176, p. 108 293, Aug. 2023, ISSN: 0098-1354. DOI: 10.1016/j.compchemeng.2023.108293.

[118] D. M. Makarov, Y. A. Fadeeva, L. E. Shmukler, and I. V. Tetko, "Machine learning models for phase transition and decomposition temperature of ionic liquids," *Journal of Molecular Liquids*, vol. 366, p. 120 247, Nov. 2022, ISSN: 0167-7322. DOI: 10.1016/j.molliq.2022.120247.

[119] G. Ren, A. M. Mroz, F. Philippi, T. Welton, and K. E. Jelfs, "Expanding the chemical space of ionic liquids using conditional variational autoencoders," *Chemical Science*, vol. 17, no. 9, pp. 4621–4631, 2026, ISSN: 2041-6539. DOI: 10.1039/d5sc08673f.

[120] T. Liu, D. Fan, Y. Chen, *et al.*, "Prediction of co2 solubility in ionic liquids via convolutional autoencoder based on molecular structure encoding," *AIChE Journal*, vol. 69, no. 10, Jul. 2023, ISSN: 1547-5905. DOI: 10.1002/aic.18182.

[121] X. Liu, J. Chu, Z. Zhang, and M. He, "Data-driven multi-objective molecular design of ionic liquid with high generation efficiency on small dataset," *Materials & Design*, vol. 220, p. 110 888, Aug. 2022, ISSN: 0264-1275. DOI: 10.1016/j.matdes.2022.110888.

[122] J. Xia, L. Zhang, X. Zhu, *et al.*, "Understanding the limitations of deep models for molecular property prediction: Insights and solutions," *Advances in Neural Information Processing Systems*, vol. 36, pp. 64 774–64 792, 2023.

[123] F. H. Vermeire and W. H. Green, "Transfer learning for solvation free energies: From quantum chemistry to experiments," *Chemical Engineering Journal*, vol. 418, p. 129 307, 2021.

[124] C. Finn, P. Abbeel, and S. Levine, "Model-agnostic meta-learning for fast adaptation of deep networks," in *International conference on machine learning*, PMLR, 2017, pp. 1126–1135.

[125] A. Nichol, J. Achiam, and J. Schulman, *On first-order meta-learning algorithms*, 2018. DOI: 10.48550/ARXIV.1803.02999. [Online]. Available: https://arxiv.org/abs/1803.02999.

[126] A. Pappu and B. Paige, "Making graph neural networks worth it for low-data molecular machine learning," *arXiv preprint arXiv:2011.12203*, 2020.

[127] C. Q. Nguyen, C. Kreatsoulas, and K. M. Branson, "Meta-learning gnn initializations for low-resource molecular property prediction," *arXiv preprint arXiv:2003.05996*, 2020.

[128] B. Dou, Z. Zhu, E. Merkurjev, *et al.*, "Machine learning methods for small data challenges in molecular science," *Chemical Reviews*, vol. 123, no. 13, pp. 8736–8780, 2023.

[129] Z. Guo, C. Zhang, W. Yu, *et al.*, "Few-shot graph learning for molecular property prediction," in *Proceedings of the web conference 2021*, 2021, pp. 2559–2567.

[130] W. Ju, Z. Liu, Y. Qin, *et al.*, "Few-shot molecular property prediction via hierarchically structured learning on relation graphs," *Neural Networks*, vol. 163, pp. 122–131, 2023.

[131] T. Schlender, M. Viljanen, J. N. van Rijn, *et al.*, "The bigger fish: A comparison of meta-learning qsar models on low-resourced aquatic toxicity regression tasks," *Environmental Science & Technology*, vol. 57, no. 46, pp. 17 818–17 830, Jun. 2023, ISSN: 1520-5851. DOI: 10.1021/acs.est.3c00334.

[132] Q. Lv, G. Chen, Z. Yang, W. Zhong, and C. Y.-C. Chen, "Meta learning with graph attention networks for low-data drug discovery," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 35, no. 8, pp. 11 218–11 230, 2024. DOI: 10.1109/TNNLS.2023.3250324.

[133] Q. Lv, J. Zhou, Z. Yang, H. He, and C. Y.-C. Chen, "3d graph neural network with few-shot learning for predicting drug–drug interactions in scaffold-based cold start scenario," *Neural Networks*, vol. 165, pp. 94–105, 2023.

[134] X. Qian, B. Ju, P. Shen, K. Yang, L. Li, and Q. Liu, "Meta learning with attention based fp-gnns for few-shot molecular property prediction," *ACS omega*, vol. 9, no. 22, pp. 23 940–23 948, 2024.

[135] Z. Meng, Y. Li, P. Zhao, Y. Yu, and I. King, "Meta-learning with motif-based task augmentation for few-shot molecular property prediction," in *Proceedings of the 2023 SIAM International Conference on Data Mining (SDM)*, SIAM, 2023, pp. 811–819.

[136] A. Kötter, S. Allenspach, C. Grebner, *et al.*, "Task-similarity is a crucial factor for few-shot meta-learning of structure-activity relationships," *ChemBioChem*, vol. 25, no. 19, e202400095, 2024.

[137] J. Eichenlaub, K. Baran, M. Śmiechowski, and A. Kloskowski, "Free volume in physical absorption of carbon dioxide in ionic liquids: Molecular dynamics supported modeling," *Separation and Purification Technology*, vol. 313, p. 123 464, 2023.

[138] K. Baran, B. Barczak, and A. Kloskowski, "Modeling lignin extraction with ionic liquids using machine learning approach," *Science of The Total Environment*, vol. 935, p. 173 234, Jul. 2024, ISSN: 0048-9697. DOI: 10.1016/j.scitotenv.2024.173234.

[139] K. Paduszyński, "Extensive databases and group contribution qsprs of ionic liquids properties. 2. viscosity," *Industrial & Engineering Chemistry Research*, vol. 58, no. 36, pp. 17 049–17 066, Aug. 2019, ISSN: 1520-5045. DOI: 10.1021/acs.iecr.9b03150.

[140] K. Paduszynski, "Extensive databases and group contribution qsprs of ionic liquid properties. 3: Surface tension," *Industrial & Engineering Chemistry Research*, vol. 60, no. 15, pp. 5705–5720, Apr. 2021, ISSN: 1520-5045. DOI: 10.1021/acs.iecr.1c00783.

[141] A. Klamt, "The cosmo and cosmo-rs solvation models," *Wiley Interdisciplinary Reviews: Computational Molecular Science*, vol. 1, no. 5, pp. 699–709, 2011.

[142] G. Landrum, P. Tosco, B. Kelley, *et al.*, *Rdkit/rdkit: 2025-03-4 (q1 2025) release*, 2025. DOI: 10.5281/ZENODO.591637. [Online]. Available: https://zenodo.org/doi/10.5281/zenodo.591637.

[143] A. Banerjee and K. Roy, "Arka: A framework of dimensionality reduction for machine-learning classification modeling, risk assessment, and data gap-filling of sparse environmental toxicity data," *Environmental Science: Processes & Impacts*, vol. 26, no. 6, pp. 991–1007, 2024.

[144] J. Demšar, T. Curk, A. Erjavec, *et al.*, "Orange: Data mining toolbox in python," *Journal of Machine Learning Research*, vol. 14, pp. 2349–2353, 2013. [Online]. Available: http://jmlr.org/papers/v14/demsar13a.html.

[145] S. M. Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions," in Curran Associates, Inc., 2017. [Online]. Available: http://papers.nips.cc/paper/7062-a-unified-approach-to-interpreting-model-predictions.pdf.

[146] M. T. Ribeiro, S. Singh, and C. Guestrin, ""why should I trust you?": Explaining the predictions of any classifier," in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Francisco, CA, USA, August 13-17, 2016*, 2016, pp. 1135–1144.

[147] E. O. K. L. et al. "Eli5" a python package which helps to debug machine learning classifiers and explain their predictions." (), [Online]. Available: https://github.com/eli5-org/eli5.

[148] N. M. O'Boyle, M. Banck, C. A. James, C. Morley, T. Vandermeersch, and G. R. Hutchison, "Open babel: An open chemical toolbox," *Journal of Cheminformatics*, vol. 3, no. 1, Oct. 2011, ISSN: 1758-2946. DOI: 10.1186/1758-2946-3-33.

[149] P. Reiser, M. Neubert, A. Eberhard, *et al.*, "Graph neural networks for materials science and chemistry," *Communications Materials*, vol. 3, no. 1, p. 93, 2022.

[150] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," in *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers)*, 2019, pp. 4171–4186.

[151] J. Demšar, G. Leban, and B. Zupan, "Freeviz—an intelligent multivariate visualization approach to explorative analysis of biomedical data," *Journal of biomedical informatics*, vol. 40, no. 6, pp. 661–671, 2007.

[152] L. Van der Maaten and G. Hinton, "Visualizing data using t-sne.," *Journal of machine learning research*, vol. 9, no. 11, 2008.

[153] J. Eichenlaub, K. Baran, K. Urbański, *et al.*, "In search of the perfect composite material—a chemoinformatics approach towards the easier handling of dental materials," *International Journal of Molecular Sciences*, vol. 26, no. 17, p. 8283, Aug. 2025, ISSN: 1422-0067. DOI: 10.3390/ijms26178283.

[154] C. De Mol, E. De Vito, and L. Rosasco, "Elastic-net regularization in learning theory," *Journal of Complexity*, vol. 25, no. 2, pp. 201–230, Apr. 2009, ISSN: 0885-064X. DOI: 10.1016/j.jco.2009.01.002.

[155] J. H. Friedman, "Greedy function approximation: A gradient boosting machine," *Annals of statistics*, pp. 1189–1232, 2001.

[156] G. E. Hinton, "Learning multiple layers of representation," *Trends in cognitive sciences*, vol. 11, no. 10, pp. 428–434, 2007.

[157] D. E. Rumelhart, G. E. Hinton, and R. J. Williams, "Learning representations by back-propagating errors," *nature*, vol. 323, no. 6088, pp. 533–536, 1986.

[158] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *International Conference on Learning Representations*, Originally published as arXiv:1412.6980, 2015.

[159] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *nature*, vol. 521, no. 7553, pp. 436–444, 2015.

[160] A. Vaswani, N. Shazeer, N. Parmar, *et al.*, "Attention is all you need," *Advances in neural information processing systems*, vol. 30, 2017.

[161] W. Ahmad, E. Simon, S. Chithrananda, G. Grand, and B. Ramsundar, "Chemberta-2: Towards chemical foundation models," *arXiv preprint arXiv:2209.01712*, 2022.

[162] A. Sperduti and A. Starita, "Supervised neural networks for the classification of structures," *IEEE transactions on neural networks*, vol. 8, no. 3, pp. 714–735, 1997.

[163] T. N. Kipf and M. Welling, "Semi-supervised classification with graph convolutional networks," *arXiv preprint arXiv:1609.02907*, 2016.

[164] Python Core Team, *Python: A dynamic, open source programming language*, Python Software Foundation, 2019. [Online]. Available: https://www.python.org/.

[165] M. Fey and J. E. Lenssen, "Fast graph representation learning with pytorch geometric," *arXiv preprint arXiv:1903.02428*, 2019.

[166] T. pandas development team, *Pandas-dev/pandas: Pandas*, version latest, Feb. 2020. DOI: 10.5281/zenodo.3509134.

[167] F. Pedregosa, G. Varoquaux, A. Gramfort, *et al.*, "Scikit-learn: Machine learning in Python," *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.

[168] T. Chen, T. He, M. Benesty, *et al.*, *Xgboost: Extreme gradient boosting*, R package version 3.3.0.0, 2026. [Online]. Available: https://github.com/dmlc/xgboost.

[169] Y. Shi, G. Ke, D. Soukhavong, *et al.*, *Lightgbm: Light gradient boosting machine*, R package version 4.6.0, 2025. [Online]. Available: https://github.com/Microsoft/LightGBM.

[170] H. Moriwaki, Y.-S. Tian, N. Kawashita, and T. Takagi, "Mordred: A molecular descriptor calculator," *Journal of Cheminformatics*, vol. 10, no. 1, Feb. 2018, ISSN: 1758-2946. DOI: 10.1186/s13321-018-0258-y.

[171] N. V. Thieu, *Permetrics: A framework of performance metrics for machine learning models*, 2024. DOI: 10.5281/ZENODO.3951205.

[172] J. D. Hunter, "Matplotlib: A 2d graphics environment," *Computing in Science & Engineering*, vol. 9, no. 3, pp. 90–95, 2007. DOI: 10.1109/MCSE.2007.55.

[173] M. L. Waskom, "Seaborn: Statistical data visualization," *Journal of Open Source Software*, vol. 6, no. 60, p. 3021, 2021. DOI: 10.21105/joss.03021. [Online]. Available: https://doi.org/10.21105/joss.03021.

[174] B. Ramsundar, P. Eastman, P. Walters, V. Pande, K. Leswing, and Z. Wu, *Deep Learning for the Life Sciences*. O'Reilly Media, 2019, https://www.amazon.com/Deep-Learning-Life-Sciences-Microscopy/dp/1492039837.

[175] S. M. Arnold, P. Mahajan, D. Datta, I. Bunner, and K. S. Zarkias, "Learn2learn: A library for meta-learning research," *arXiv preprint arXiv:2008.12284*, 2020.