Universität des Saarlandes

Master Thesis

Enumeration-aware Molecular SMILES Transformers for Representation Learning and Low-resource Scenarios

submitted in fulfillment of the degree requirements of the

MSc in Computer Science at Saarland University

Author:

Shahrukh Khan

Matriculation: 7004431

Supervisors:

Prof. Dr. Dietrich Klakow

Prof. Dr. Olga Kalinina

April 8, 2023

Contents

Eidesstattliche Erklärung

Ich erkläre hiermit an Eides Statt, dass ich die vorliegende Arbeit selbstständig verfasst und keine anderen als die angegebenen Quellen und Hilfsmittel verwendet habe.

Statement in Lieu of an Oath

I hereby confirm that I have written this thesis on my own and that I have not used any other media or materials than the ones referred to in this thesis.

Einverständniserklärung

Ich bin damit einverstanden, dass meine (bestandene) Arbeit in beiden Versionen in die Bibliothek der Informatik aufgenommen und damit veröffentlicht wird.

Declaration of Consent

I agree to make both versions of my thesis (with a passing grade) accessible to the public by having them added to the library of the Computer Science Department.

Saarbrücken, April 8, 2023

Shahrukh Khan

Acknowledgements

This thesis is realized thanks to the helpful support and guidance of multiple people. I am grateful to my supervisors Prof. Dr. Dietrich Klakow and Prof. Dr. Olga Kalinina for allowing me to partake in this research endeavor under their supervision. I would also like to thank Prof. Dr. Andrea Volkamer for helping me distill the domain-specific cheminformatics knowledge related to my research. Moreover, I would also like to sincerely thank all members of the NextAID group for providing their useful feedback, which allowed me to make meaningful refinements to this thesis. Finally, I would also like to appreciate Awantee Deshpande [Deshpande, 2022] and Hassan Soliman [Soliman, 2022] for their well-structured thesis documents which inspired me to enhance the quality of my thesis.

On a personal note, I would like to express my deepest gratitude to my family and friends for pledging their emotional support and prayers for me. I am eternally grateful to my parents for being my persistent source of encouragement through all phases of my life. I would also like to wholeheartedly thank my wife Rahat for her patience and support during my graduate studies. Also, I would like to genuinely appreciate my brother Faisal, who has always been there to help me reprieve during tough times. Lastly, I would like to convey my heartiest gratitude to all my friends here and in Pakistan for their everlasting prayers and for motivating me throughout my Master's.

Abstract

Computer-aided drug discovery plays a pivotal role in the pipeline of discovering novel drugs. Notably, intelligent machine learning techniques have made drug development pipelines even more efficient in the pharmaceutical domain. In addition to discovering novel drugs, machine learning methods also help practitioners establish critical properties (i.e. toxicity) of the molecules that constitute a drug. Recent work has primarily been based on using deep transfer learning methods to learn meaningful molecular fingerprints. Importantly, the eventual success of these techniques heavily relies on the fidelity and richness of the learned fingerprints.

One of these paradigms deals with learning molecular fingerprints from SMILES representations in a self-supervised manner without the need for them to be labeled. It entails pre-training neural language models with objectives including token masking on SMILES and multi-task regression on the physicochemical properties of molecules. Specifically, in recent work, neural fingerprints learned with BERT-like language models have significantly outperformed classical machine learning methods on drug-discovery-related tasks including *Quantitative Structure-Activity Relationship* and *Virtual Screening*. However, current SMILES molecular pre-training regimes underperform in low-data settings. This is potentially attributed to the low fidelity of the learned fingerprints due to the absence of enumeration knowledge.

In this thesis, we address this challenge in a two-faceted manner: (1) by introducing novel SMILES pre-training objectives based on contrastive learning, and denoising respectively to incorporate SMILES enumeration knowledge into learned fingerprints, (2) by coupling transfer learning with semi-supervised learning approaches to adapt to small-data settings. Our experimental results demonstrate that making pre-trained molecular language models enumeration-aware with contrastive learning enhances their performance on downstream tasks. Precisely, as much as 9% improvement on the AUROC metric for the *Quantitative Structure-Activity Relationship* task and over 10% and 5% performance gain on the AUROC and BEDROC20 metrics respectively for the *Virtual Screening* task.

Furthermore, we extend our work on transfer learning to low-data scenarios suffering from label scarcity. Our empirical results on the *MoleculeNet Benchmark* show that replacing fully supervised fine-tuning with semi-supervised learning can yield up to 11% improvement in AUROC score on the test datasets including Tox21 and BACE while training on a small labeled dataset. Similarly, we apply our proposed approaches on an actual low-resource molecular dataset by Helmholtz Institute for Pharmaceutical Research Saarland and observe a gain of $\sim$3% on the AUROC metric. Hence showing, that using semi-supervised for fine-tuning can effectively mitigate the scarcity of labels in molecular datasets while employing pre-trained language models.

Keywords: Machine Learning, Drug Discovery, Quantitative Structure-Activity Relationship, Virtual Screening, Masked Language Modelling, Semi-supervised Learning, Domain Adaptation, Masked Token Prediction, Contrastive Learning, SMILES Denoising

Chapter 1
Introduction

1.1 Motivation

Deep Neural Networks have recently made groundbreaking breakthroughs in the field of *Machine Learning* (ML). More precisely, with the introduction of *Self-supervised Learning* (SSL) deep learning techniques have even surpassed human performance in the *Natural Language Processing* (NLP) [Devlin et al., 2019] and the domain of computer vision domain [Kolesnikov et al., 2021]. SSL entails ML model learning from a large unlabeled dataset whilst generating labeled targets itself autonomously. Consequently, this eradicates the fundamental bottleneck of having a large labeled training dataset prior to training a generalizable ML model. Importantly, this opens a corridor of opportunity for optimizing the drug discovery process, which usually spans between 6 and 10 years [Mohs and Greig, 2017].

Graph-based ML approaches have traditionally surpassed SMILES-based language models on drug discovery tasks such as *Quantitative Structure-Activity Relationship* (QSAR), and *Virtual Screening*, and *Quantitative Structure-Property Relationship* (QSPR) [Duvenaud et al., 2015, Kearnes et al., 2016]. However, such methods employ large labeled datasets. Resultantly, the scarcity of labeled data for a particular domain constrains their application. Hence, obtaining big datasets with domain-specific protein affinities and molecular properties might not be feasible for the effective application of graph-based methods.

Pre-training regimes based on SSL mainly *Masked Language Modeling* (MLM) have been employed to train SMILES-based language models [Chithrananda et al., 2020, Ahmad et al., 2022] in order to mitigate the need for large labeled datasets. This has been also made viable due to the availability of large chemical databases such as GuacaMol [Brown et al., 2019a], and ZINC20 [Irwin et al., 2020] etc. The resulting language models have demonstrated state-of-the-art results for drug discovery tasks such as QSAR, and *Virtual Screening* [Fabian et al., 2020].

Albeit MLM-based language models for molecular SMILES have proven to be performant, the majority of the existing methods do not encode the knowledge about SMILES enumerations explicitly. The absence of enumeration knowledge affects the language models in two unique ways.

Firstly, the absence of enumeration knowledge results in directly degrading the quality of representations learned by such language models. For instance, vector representations of two enumerations of the same molecule are highly likely to be distant in that particular latent space, as opposed to being close. Similarly, molecular low-quality representations have an adverse impact on performance on downstream tasks like QSAR, and *Virtual Screening* [Vamathevan et al., 2019].

Secondly, canonicalization algorithms aid in establishing distinct representations of molecules. Resultantly, it removes ambiguous artifacts from SMILES. However, canonicalization rules are prone to be intrusive in the language model's learning process. Therefore, representations learned by language models on canonical SMILES have lower fidelity [Fabian et al., 2020]. Additionally, previous studies have demonstrated that utilizing enumerations in the training dataset enhances the expressiveness of the ML model [Bjerrum and Sattarov, 2018a]. Consequently, embedding enumeration knowledge into SSL-based language models is quintessential for the expressiveness and high fidelity of the learned representations.

Another gap in the current cheminformatics literature is the lack of attention given to evaluating pre-training-based SSL approaches on low-resource (data) settings [Honda et al., 2019]. Hence, established benchmark datasets and performance metrics for evaluating cheminformatics models trained on small datasets are scarce. We intend to address such limitation by employing *Semi-supervised Learning* (SESL).

Thereby, for the effective application of pre-trained language models on low-data molecular datasets, there is a need for more domain-specific data-efficient approaches.

1.2 Limitations of State of the Art

The absence of explicitly administered enumeration knowledge into language models contributes to performance degradation on downstream tasks [Payne et al., 2020]. One potential solution was proposed by introducing binary classification as an auxiliary pre-training task [Fabian et al., 2020]. The classification task involved classifying a pair composed of canonical and enumerated SMILES as the same or different. However, the classification task interfered with primary pre-training tasks and resulted in performance degradation on QSAR tasks.

1.3 Goals of Thesis

The main objective of our work is to enable the application of pre-trained language models on low-data drug discovery tasks. Our work addresses this in a bi-fold manner, which includes the administration of domain-specific knowledge into latent representations of general-purpose pre-trained molecular language models. Furthermore, we evaluate the application of SESL methods on low-data downstream drug discovery datasets, leveraging additional in-domain unlabeled data to circumvent the scarcity of labels. Finally, we also evaluate the proposed methods on a real-world low-resource application dataset to further verify the appositeness of our work.

Concretely, this thesis primarily focuses on addressing two challenges. Those include (1) explicitly making language models enumeration-aware and (2) adapting *Bidirectional Encoder Representations from Transformers* (BERT)-like transformer models to low-resource molecular datasets. Thereby, has the following goals:

- Analysis of contrastive learning as an intermediate pre-training objective for encoding enumeration knowledge into SMILES-based language models.
- Comparison of enumeration-aware encoder-based BERT-like language models against seq2seq transformers-based language models on drug discovery downstream tasks, including QSAR and *Virtual Screening*.
- Evaluation of dense vector representations learned by enumeration-aware transformers, demonstrating contrastive learning-oriented pre-training, is quintessential to high-fidelity embeddings.
- Examination of various embedding approaches including the ones from language models, RDKit[1] physicochemical properties, fingerprints, and hybrid approaches for in-domain sample screening in low-resource scenarios.

[1] `https://www.rdkit.org/`.

- Comparison and application of various SESL techniques in tackling low-resource drug discovery tasks.

1.4 Research Questions

Application of neural approaches in low-data drug discovery is challenging amidst non-trivial constraints of lack of labeled data and availability of finite data augmentation techniques. Therefore, this thesis alleviates such challenges by addressing the following research questions:

RQ1: Can contrastive domain adaptation encode additional knowledge into encoder-based language models?

RQ2: Can we incorporate enumerations by a learning encoder-decoder-based canonicalization model?

RQ3: How does our domain adaptation pre-training compare to comparative existing state-of-the-art pre-training approaches for molecular SMILES?

RQ4: Which of the enumeration-aware pre-training approach better incorporates enumeration knowledge in representations?

RQ5: To what extent can the *Semi-supervised Learning* (SESL) paradigm help in low-resource molecular scenarios?

1.5 Structure of the Thesis

This master thesis is structured as follows: The second chapter provides a brief overview of computer-aided drug discovery. There we also introduce the essential domain-specific cheminformatics terminologies relevant to the context of our work. Furthermore, we also discuss how machine learning can be applied in the drug discovery life cycle, accompanied by prominent neural approaches.

In Chapter 3, we go over into deeper detail about the recent transfer learning approaches for molecular language models and the role of semi-supervised learning in low-data scenarios. In Chapter 4, we detail the specifics of the molecular datasets we use to evaluate our proposed approaches. We also introduce a real-world dataset that we use to evaluate our proposed approaches in a more realistic setting. Finally, we also briefly discuss the dataset-splitting approaches for molecular datasets.

In Chapter 5, we explain our proposed approaches for molecular representation learning and the application of pre-trained molecular language models in low-data settings.

In Chapter 6, we provide the necessary details about the experimental setup of our work, including the evaluation metrics corresponding to different downstream tasks. Furthermore, we also discuss the critical design choices when dealing with low-resource molecular datasets including the choice of data augmentation, and best practices for intermediate domain adaptation. In Chapter 7, we summarize the conclusions from our proposed approaches in this Master thesis and discuss potential avenues for future research.

Chapter 2
Background

Computational methods have significantly boosted the efficiency of drug discovery and development pipelines in the pharmaceutical domain [Smith et al., 2018]. Computer-aided drug discovery essentially leverages established structural knowledge of the target compounds to identify promising drugs. Furthermore, computational methods also play a pivotal role in predicting molecular properties and are quintessential to effective drug discovery. In this chapter, we provide a comprehensive overview of fundamental building blocks for understanding and applying computational methods for drug discovery. Furthermore, we also discuss the principal role of ML methods responsible for the recent advancements in drug development.

In Sec. 2.1, we present how molecules can be represented digitally, alongside the data structures that store and process them. Sec. 2.2 discusses the viability of the ML methods in the context of drug discovery. Furthermore, we also provide a brief overview of downstream molecular machine learning tasks that are relevant to our proposed methods. In Sec. 2.3, we survey the state-of-the-art sub-field of ML, which is deep learning. Particularly, we broadly study the key classes of deep learning architectures and paradigms such as transfer learning, which make it one of the most successful approaches for solving non-trivial drug discovery tasks.

2.1 Computational Representation of Molecules

Before delving into any computational methods for processing molecules, it is important to answer the question, how molecules can be represented digitally? Primarily, molecules can be represented as graphs, SMILES, 3D structures, and vector-based fingerprints

[Cordero, 2021]. These representations are visually illustrated in Fig. 2.1.

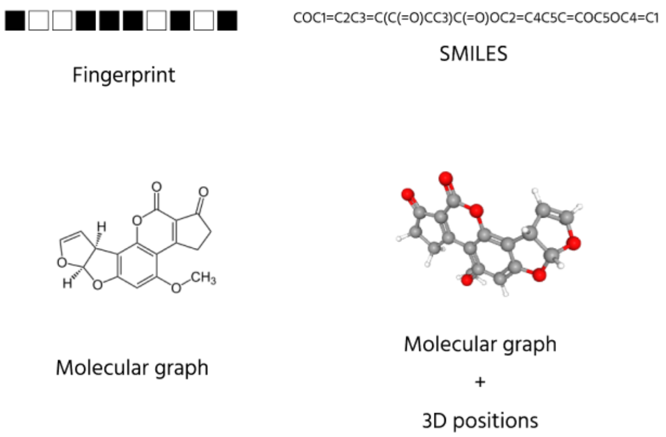

Figure 2.1: Computational representations of molecules by [Cordero, 2021]

.

2.1.1 Molecular Graph

A molecular graph consists of nodes that represent the atoms and edges which represent the bonds between the atoms. There can be multiple properties associated with an atom such as its atomic number, atomic weight, charge, and the number of hydrogen atoms attached to it. Whereby, the properties of bonds are more trivial. Specifically, a bond can be of type single, double, triple, or aromatic. The aromatic property refers to a set of rules that are part of the chemical nomenclature system [James, 2004].

Computationally, molecular graphs are represented as adjacency matrices. Importantly, the graphs can either be directly processed by the downstream computational algorithms, unlike SMILES which first needs to be in a vector notation. Furthermore, graph-based ML techniques can learn fingerprints while encoding the graph properties of the molecules into the learned fingerprints.

2.1.2 SMILES

With *Simplified Molecular-Input Line-Entry System* (SMILES) molecules are represented as strings [Weininger, 1988]. Furthermore, in contrast to the graph representation, the SMILES format can be viewed as a simple language with few grammatical rules. Importantly, SMILES representation holds the commutative property with respect to the arrangement of atoms within a molecule. Hence, two important transformations are applicable to SMILES which are discussed below:

Enumeration

A single chemical molecule can have multiple corresponding SMILES strings also known as Enumerations. Enumerations are achieved by random permutation of the atoms present in the molecule [Bjerrum, 2017]. Fig. 2.2 shows a visual example of the application of enumeration on the toluene molecule.

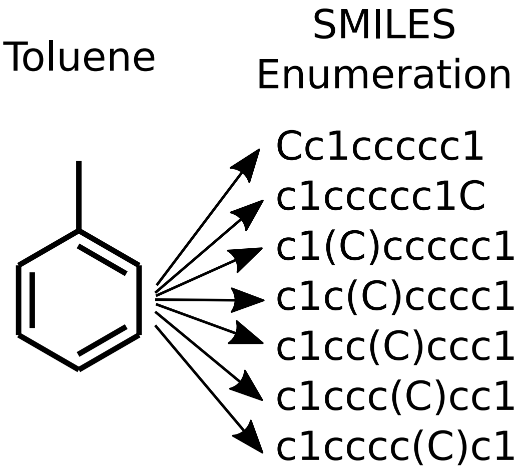

Figure 2.2: An exemplary illustration of enumeration of the toluene molecule by [Bjerrum, 2017].

Canonicalization

The canonicalization technique was introduced for the purpose of molecular disambiguation [Weininger et al., 1989]. It primarily establishes a one-to-one correspondence between molecules and their corresponding SMILES representations.

It is important to note here that we use SMILES representations for the implementation and evaluation of our proposed approaches in this thesis. Since our work primarily centers around language models which only deal with the textual representation of data.

2.1.3 Fingerprint

Molecular fingerprints encode the molecule structure as vectors of numbers. The most pervasive form of molecular fingerprints is binary (bits) vectors indicating the presence (represented as 1s) or absence (represented as 0s) of a specific molecular substructure. These fingerprints are useful in various drug discovery tasks such as computing the similarity between two molecules, querying active molecules from a large database comprising numerous decoys and actives, etc. Importantly, recent neural machine learning approaches can help learn more robust continuous molecular fingerprints from data in an unsupervised fashion. Further details about learning molecular fingerprints are presented in Sec. 2.3.

2.1.4 3D Structure

The 3D molecular structure is also a molecular graph, albeit in a three-dimensional space. The three-dimensional structure encodes additional information about the shape of the molecule. Moreover, the shape in the three-dimensional space of the molecule is highly reliant on the spatial orientation of the bonds of the molecule [Reusch, 1999].

2.2 Drug Discovery Methods for Machine Learning

Machine learning has become an integral part of modern computer-aided drug discovery and design pipelines [Klambauer et al., 2019]. The applications of ML in drug discovery include but are not limited to *Virtual Screening*, QSAR tasks, QSPR studies, and de novo drug architectures. In this section, we explain *Virtual Screening* and QSAR tasks which are quintessential to the evaluation of our proposed methods as described in detail in Chapter 6.

2.2.1 QSAR

The goals of QSAR studies are twofold. First, they help establish a relationship between chemical structures and biological activities based on a molecular dataset. Second, to predict the activities of novel molecules. Hence, standard ML approaches based on classification and regression are appropriate to develop QSAR models. [2] Here, the predictors are either fingerprints from molecular structures or molecular physicochemical properties. The physicochemical properties are measurable physical properties of chemical molecules based on different molecular attributes of that particular molecule. Whereby, the response variables are based on at least one biological activity such as health toxicity or ecotoxicity.

2.2.2 Virtual Screening

Virtual screening corresponds to a cheminformatics approach for retrieving a few active molecules among numerous decoys. Here, the active molecules are those which are likely to bind to a drug target. Moreover, the active and the decoys are specific to that drug target, such as a protein or enzyme. Furthermore, the retrieval is typically performed by computing a similarity measure i.e. cosine, etc., over molecular fingerprints. Fig. 2.3

[2] https://en.wikipedia.org/wiki/Quantitative_structure%E2%80%93activity_relationship.

shows a schematic demonstration of *Virtual Screening*. Importantly, the performance of the *Virtual Screening* task is highly dependent on the richness and fidelity of the learned or computed fingerprints. In our work, we introduce novel CL-based multi-task neural approaches (as described in Chapter 5) for learning high-fidelity fingerprints.

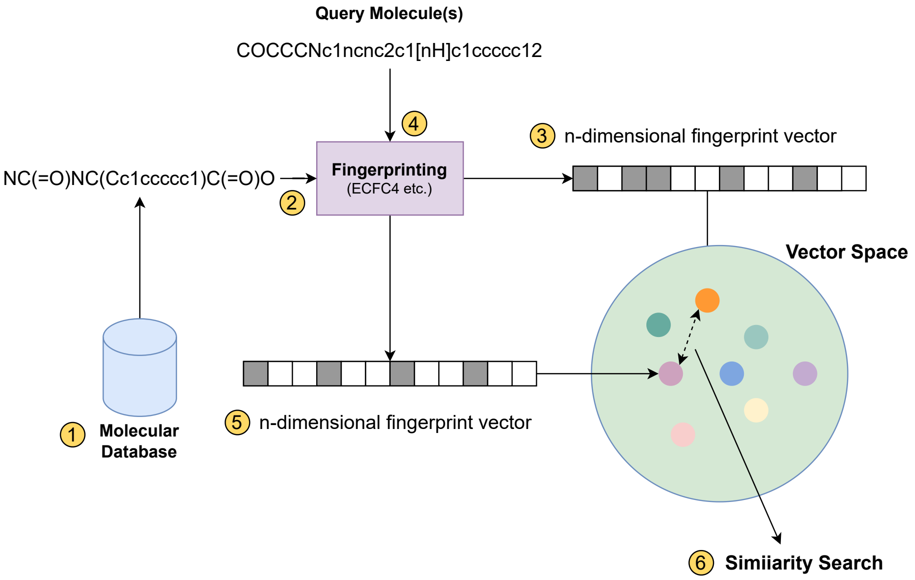
Query Molecule(s)
COCCCNc1ncnc2c1[nH]c1ccccc12
4
3 n-dimensional fingerprint vector
NC(=O)NC(Cc1ccccc1)C(=O)O
2
Fingerprinting
(ECFC4 etc.)
Vector Space
1 Molecular Database
5 n-dimensional fingerprint vector
6 Simiiarity Search

Figure 2.3: An overview of an exemplary molecular *Virtual Screening* pipeline.

2.3 Neural Transfer Learning for Drug Discovery

Traditional machine learning approaches for downstream drug discovery tasks i.e., virtual screening, QSAR, etc. rely on sparse rule-based fingerprints as predictors. However, recent state-of-the-art machine learning approaches are predominantly based on its special class of algorithms called neural networks. The key differentiator of neural approaches is their underlying ability to learn dense, high-fidelity and rich fingerprints from the molecular dataset. As a result, these learned fingerprints self-contain additional contextual information from the molecular structure and composition.

Specifically, most of the recent successful fingerprint-learning approaches have been based on the Transformer architecture [Vaswani et al., 2017]. The key ingredient behind the powerful transformer architecture is *Multi-head Attention* mechanism, which effectively encodes multi-faceted context while learning fingerprints as shown in Fig. 2.4. Consequently, several architectures based on the transformer model have been proposed incorporating the paradigm of **transfer learning**.

Transfer learning entails a two-step training regime for the neural network. In the first stage, the network is trained typically in an unsupervised fashion on a large general domain database of SMILES, this is referred to as *Pre-training*. Whereby, in the second stage the pre-trained model is trained in a supervised fashion on a comparatively smaller molecular dataset, this process is known as fine-tuning. Objectively, such transformer-based architectures coupled with transfer learning can be categorized into two classes of architectures. These classes are namely the encoder architecture and the encoder-decoder architecture.

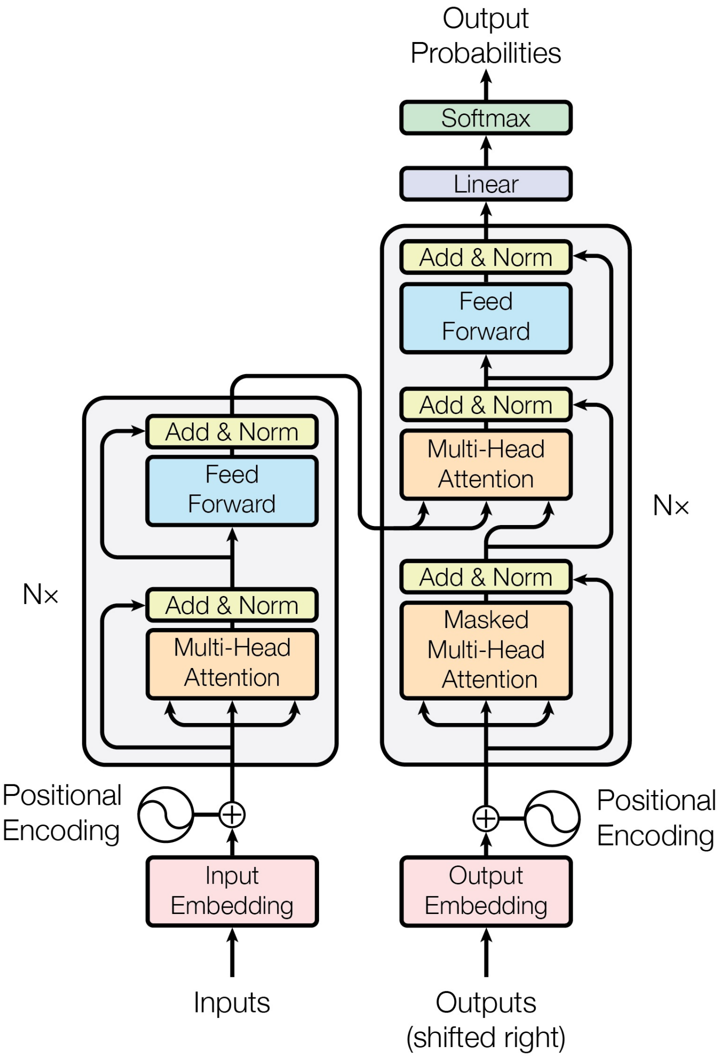

Figure 2.4: The Transformer architecture based on the Encoder and Decoder modules by [Vaswani et al., 2017]

.

2.3.1 Encoder Architecture

The BERT model is a highly performant variant of transformer architecture [Devlin et al., 2019]. As shown in Fig. 2.5, the pre-training phase of the BERT model involves either training with MLM or *Next Sentence Prediction* (NSP) objective on a large unlabeled text corpus. Subsequently, the pre-trained model is adapted in a supervised fashion on a downstream task during fine-tuning.

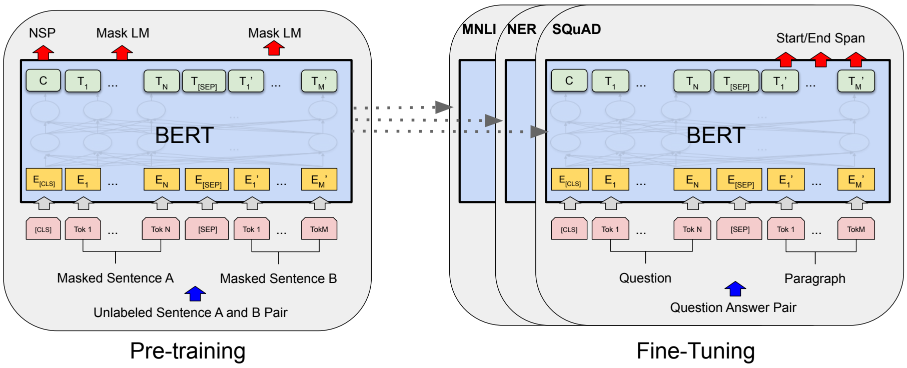

Figure 2.5: Two phase training workflow for the BERT model based on transfer learning [Devlin et al., 2019]

.

Pre-training a BERT-like model on a molecular SMILES dataset slightly differs from its counterpart pre-trained on a natural language corpus. Since molecular SMILES are analogous to single sentence constructs, hence NSP pre-training objective is not viable. Furthermore, recent work has shown alternative pre-training mechanisms based on predicting the physicochemical properties of the molecule known as *Multi-task Regression* (MTR) [Ahmad et al., 2022, Fabian et al., 2020]. MTR, unlike MLM-based pre-training, decouples the relationship between QSAR biological activities and the structure of the SMILES. Concretely, the plausible pre-training mechanisms for SMILES-based BERT-like models are MLM and MTR. Whereby, the fine-tuning stage for downstream drug discovery tasks remains synonymous with NLP.

In our work, we propose an additional pre-training objective to complement MLM-based pre-training. Our proposed pre-training objective relies on contrastive learning, which aims to make BERT-like molecular language models enumeration-aware. Further details about our proposed approaches are specified in Sec. 5.1.1.

2.3.2 Encoder-decoder Architecture

The *Bidirectional and Auto-Regressive Transformer* (BART) architecture is an encoder-decoder transformer composed of a bidirectional BERT-like encoder and an autoregressive decoder [Lewis et al., 2019]. The encoder block consists of multi-head attention blocks similar to the BERT architecture. Whereby, the autoregressive behavior of the decoder entails that it only conditions the next prediction based on the prior outputs, and hence has no prior knowledge of future tokens. A high-level architecture of the BART model is shown in Fig. 2.6.

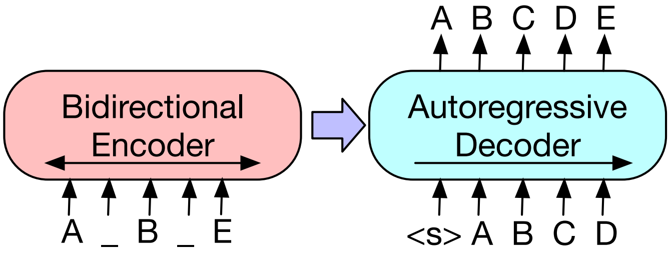

Figure 2.6: BART encoder-decoder architecture by [Lewis et al., 2019] for denoising-based pre-training

The BART model pre-training is based on corrupting the input sequences with arbitrary noising functions. Subsequently, the model is then trained with the assistance of a denoising objective, which outputs the original sequence prior to corruption. There are various plausible sequence corruption schemes employed by the authors. These noising functions include random shuffle of sentences, random deletion of tokens, document rotation, token masking, and randomly selected text span replacement with a mask token as illustrated in Fig. 2.7.

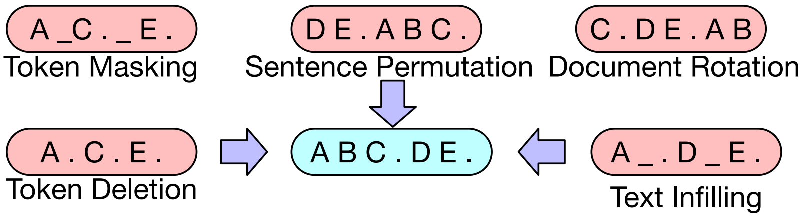

Figure 2.7: Input corruption mechanisms for BART pre-training by [Lewis et al., 2019]

Not all of the above noising schemes are applicable to molecular SMILES because of alternate grammar. Hence, we alter the noising functions for SMILES whilst BART pre-training. The details about our proposed molecular noising schemes are described in Sec. 5.1.2.

2.4 Summary

In this chapter, we described the fundamentals for applying machine learning to molecular datasets. Precisely, we first presented the digital representations of molecules, including molecular graphs and SMILES. We also discussed the important properties of SMILES representation, since our work builds on that.

Furthermore, in the context of machine learning, we presented the important downstream machine learning tasks of *QSAR* and *Virtual Screening*. Finally, we introduced the notion of transfer learning for molecular drug discovery tasks and discussed its most performant variant based on the transformer architecture. In Chapter 3, we discuss the applications of the transformer architecture for learning context-aware fingerprints for computational drug discovery tasks. Additionally, we also highlight possible ways to further refine the fingerprint learning process.

Chapter 3
Related Work

In this chapter, we present a holistic overview of various neural approaches applied to downstream drug discovery tasks. In Sec. 3.1, we discuss two prominent classes of architectures (encoder-based and encoder-decoder) for learning SMILES fingerprints in the molecular latent space. Whereby, Sec. 3.2 focuses on applications of such neural approaches in the low-data settings. Moreover, we outline strategies such as data augmentation and semi-supervised learning to mitigate the absence of further labeled data. Finally, we briefly summarize the context in which the discussed approaches are relevant to our proposed methods in Sec. 3.3.

3.1 SMILES Representation Learning

The performance of machine learning models on *Virtual Screening* and QSAR tasks is directly influenced by the fidelity of molecular representations also known as fingerprints [Vamathevan et al., 2019]. Traditionally, molecular fingerprints have been calculated using rule-based algorithms such as ECFP4 [Rogers and Hahn, 2010], and Morgan fingerprint [Morgan, 1965] etc. Rule-based fingerprints produce sparse vector representations for molecular SMILES, which are distributed over a discrete space. However, such algorithms fail to produce rich molecular representations for low-data settings.

Recent work has successfully demonstrated that molecular representations learned by language models have higher fidelity. Specifically, the neural fingerprints from language models have outperformed rule-based algorithms on tasks including molecular reaction [Liu et al., 2017, Schwaller et al., 2019] and property prediction [Jastrzebski et al., 2016, Winter et al., 2019].

As discussed in Sec. 2.3, there are two foundational neural network architectures used in molecular language models, (1) encoder models, and (2) encoder-decoder models. Thereby, in Sec. 3.1.1, and Sec. 3.1.2 we go over noteworthy recent work related to encoder-only and encoder-decoder transformer-based language models respectively. Additionally, our work introduces contrastive learning to incorporate enumeration knowledge in learned molecular fingerprints. Hence, we also survey the related work in the area of contrastive learning in Sec. 3.1.3.

3.1.1 Encoder Models

Most of the previous work on encoder-based molecular language models is inspired by the NLP literature. Thereby, the majority of the recent approaches are based on the transformer architecture [Vaswani et al., 2017]. Precisely, the BERT architecture combines transfer learning with SSL using the encoder part of the transformers [Devlin et al., 2019].

Conventionally, BERT-like models have been trained on large molecular SMILES databases with MLM objective also known as pre-training phase [Payne et al., 2020]. MLM entails masking a portion of input tokens randomly whilst the model predicts the masked tokens as output as shown in Fig. 3.1. Subsequently, the cross-entropy loss is minimized between predicted tokens and original input tokens. Consequently, this allows the language model to acquire intrinsic topological knowledge of the chemical space for better performance on property prediction tasks [Chithrananda et al., 2020]. Finally, pre-trained models are further trained in a supervised learning setting called fine-tuning. Fine-tuning is typically performed on downstream QSAR and QSPR tasks such as molecular property prediction [Wang et al., 2019, Chithrananda et al., 2020].

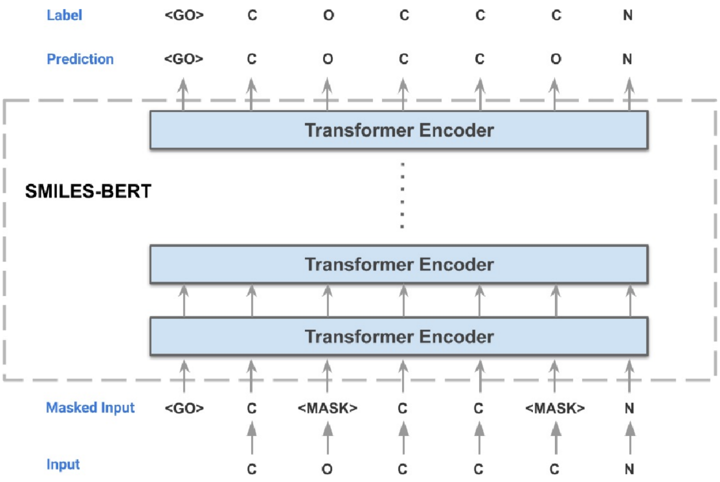

Figure 3.1: SMILES-BERT pre-training with MLM [Wang et al., 2019].

Augmenting the MLM-based BERT pre-training with auxiliary pre-training objectives has also shown promising outcomes. For instance, MolBERT [Fabian et al., 2020] uses two additional pre-training objectives in addition to MLM, as shown in Fig. 3.2. The first task implies SMILES pair classification, where the second SMILES in the pair can either be the enumeration of the first SMILES or an entirely different one. Thereafter, the model is trained by minimizing the binary cross-entropy loss. The second task involves predicting over 200 real-valued descriptors concerning physicochemical molecular properties extracted using RDKit. Here, the model learns to optimize the mean squared error over all predicted descriptors in an MTR fashion.

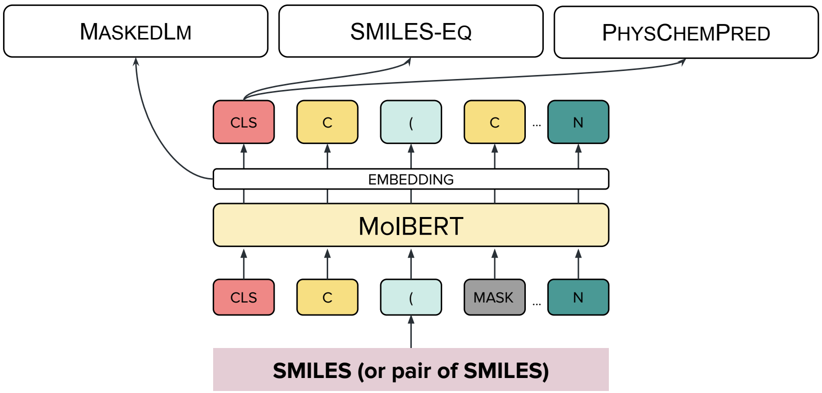

Figure 3.2: Schematic architecture of the MolBERT language model [Fabian et al., 2020].

Similarly, Chemberta-2 [Ahmad et al., 2022] independently pre-trains with either MTR objective over physicochemical molecular properties from RDKit or alternatively with MLM as shown in Fig. 3.3. Both Chemberta-2 and MolBERT then finally fine-tune the pre-trained BERT-like models on supervised QSAR tasks.

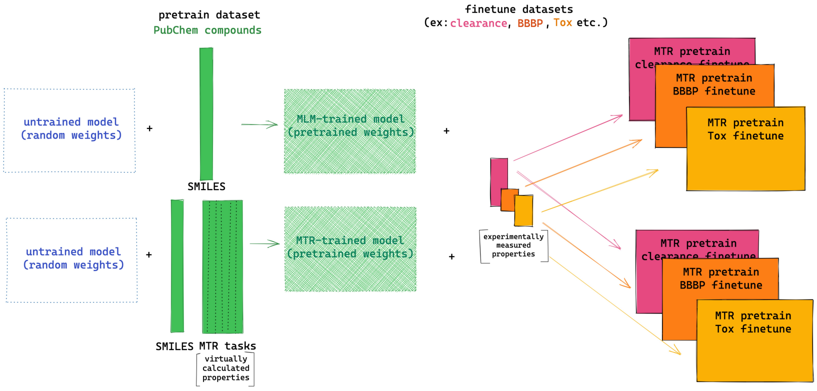

Figure 3.3: ChemBERTa-2 pre-training and fine-tuning pipeline [Ahmad et al., 2022].

Conversely, there also have been studies that proposed modifications to the transformer encoder to incorporate additional molecular knowledge into the language models. For instance, instead of pre-training on complete molecules, Mol-BERT [Li and Jiang, 2021] considers only molecular substructures. The molecular fragments are extracted from the Morgan algorithm, thereby mapping SMILES to biologically inspired words and sentences.

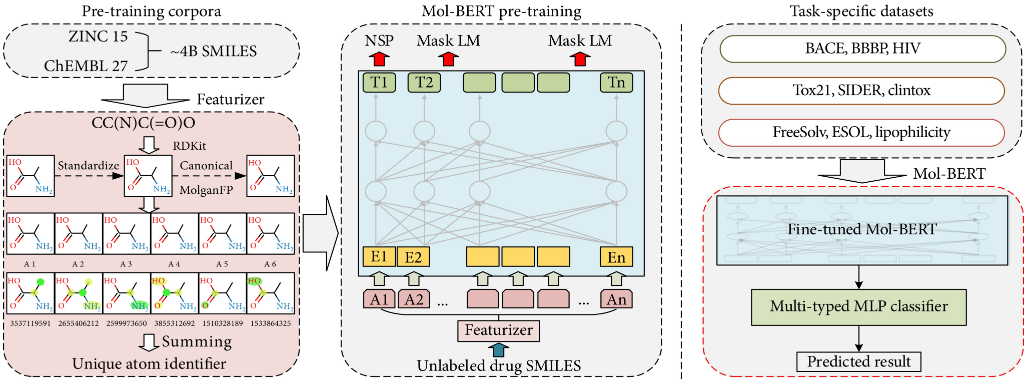

Figure 3.4: Full training pipeline for Mol-BERT [Li and Jiang, 2021].

Similarly, some work has also been done on altering the attention mechanism to include molecular graph knowledge. For example, *Molecule Attention Transformer* (MAT) [Maziarka et al., 2020] modifies the attention mechanism from the original transformers to embed inter-atomic distances and information relating to the structure of the molecular graph.

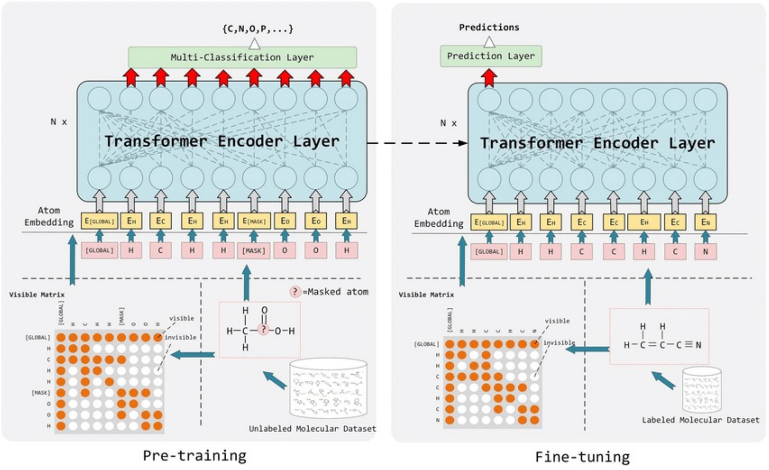

Figure 3.5: Overall training pipeline using transfer learning of the MG-BERT language model [Zhang et al., 2021a].

Identically, as shown in Fig. 3.5, the *Molecular Graph BERT* (MG-BERT) [Zhang et al., 2021a] combines the capabilities of *Graph Neural Networks* GNNs with BERT-like language models. Specifically, it amends the attention mechanism to incorporate the message-passing mechanism from GNNs. Resultantly, the MG-BERT language model can learn from molecular graphs and SMILES strings.

3.1.2 Encoder-decoder Models

The molecular SMILES-based pre-training using encoder-decoder predates the encoder-based SMILES pre-training regimes. As such, encoder-decoder have been used for pre-training on SMILES in two ways (1) using variational autoencoders [Kingma and Welling, 2013], (2) *Sequence to Sequence* (seq2seq) models. Concretely, the encoder-decoder models output the SMILES by decoding using the learned latent representations.

Autoencoders have primarily been employed in QSAR studies, such as molecular property prediction [Gómez-Bombarelli et al., 2018]. This involves a two-step transfer learning approach, as shown in Fig. 3.6. Firstly, the autoencoder is pre-trained in an unsupervised fashion. Here, the high-dimensional discrete input (SMILES) is projected onto low-dimensional continuous space also known as a bottleneck. Consequently, that low-dimensional latent representation is then used to reconstruct the discrete input as the prediction. Secondly, once the autoencoder has been pre-trained on a large unlabeled dataset, the learned latent representations are then used for a supervised downstream QSAR task [Kusner et al., 2017, Gómez-Bombarelli et al., 2018].

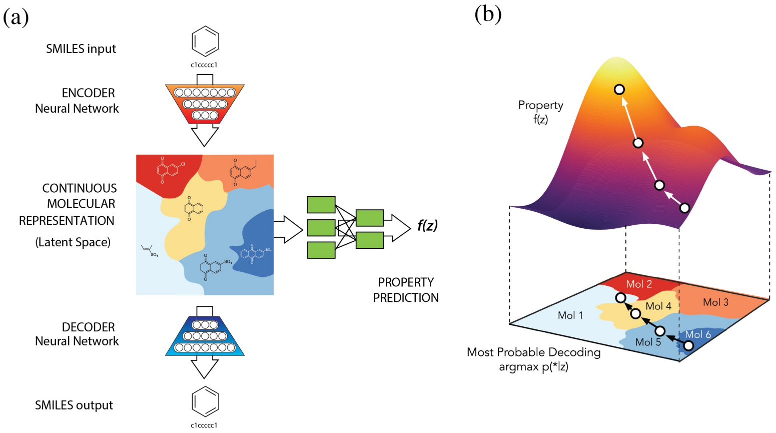

Figure 3.6: Autoencoders for fingerprint learning [Gómez-Bombarelli et al., 2018].

The vanilla molecular autoencoders have previously suffered from two challenges. Either they lacked the knowledge about SMILES enumerations or performed unconstrained decoding on SMILES [Bjerrum and Sattarov, 2018b, O'Boyle and Dalke, 2018]. To accommodate the SMILES enumerations knowledge, an autoencoder is proposed that predicts SMILES enumerations by using canonical SMILES as input [Bjerrum and Sattarov, 2018b]. This maximizes both the similarity between the latent representations and the molecular fingerprint similarity in the molecular space.

Contrary to the autoencoders which learn dense molecular representations as fingerprints only with the reconstruction objective. The seq2seq SMILES models have been trained with different pre-training objectives. These include pre-training methods such as machine translation, denoising, and input reconstruction objectives [Winter et al., 2019, Honda et al., 2019, Irwin et al., 2022, Xu et al., 2017].

Variants of sequence modeling *Recurrent Neural Networks* (RNNs) such as *Long Short-term Memory* (LSTM) [Hochreiter and Schmidhuber, 1997], and *Gated Recurrent Unit* (GRU) [Chung et al., 2014] have been used as the seq2seq SMILES models. The previous work includes training such GRU-based RNNs combined with the attention mechanism [Bahdanau et al., 2014] with SMILES reconstruction objective [Xu et al., 2017]. The attention mechanism assists the network to learn representations in a compact and centralized latent fingerprint space.

Analogous to the RNNs and the encoder-based representation learning models, the vanilla transformer architecture [Vaswani et al., 2017] has also demonstrated promising results for the SMILES reconstruction task. The transformer architecture essentially embellishes the attention mechanism with multi-head attention, allowing us to learn multifaceted properties about the molecular space. [Honda et al., 2019].

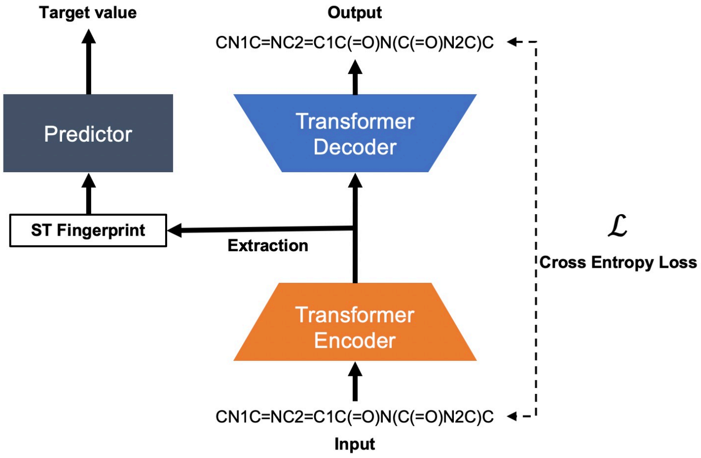

Figure 3.7: Illustration of the SMILES transformer architecture for low-data molecular scenarios [Honda et al., 2019].

Lastly, one of the previous works also highlights the tremendous capabilities of transformers-based denoising autoencoders on QSAR tasks [Irwin et al., 2022]. Precisely, it employs the BART model [Lewis et al., 2019] which has produced state-of-the-art results on the downstream QSAR tasks. The BART model takes corrupted input SMILES with noise coming from enumerations or random masking. The decoder is responsible for reconstructing the original canonical SMILES from the latent representation while removing the noise. Hence, this helps the denoising model to establish both a syntactical and semantic understanding of the molecular SMILES and the underlying chemical space. However, in this work, the authors do not evaluate the pre-trained encoder-decoder model on downstream classification-based QSAR and *Virtual Screening* tasks.

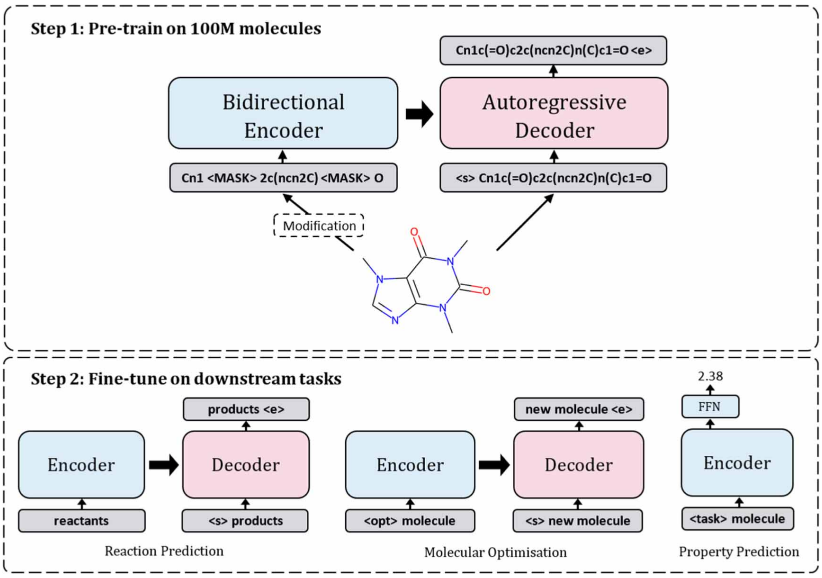

Figure 3.8: Denoising-based Chemformer pre-training pipeline for learning SMILES fingerprints [Irwin et al., 2022].

3.1.3 Contrastive Learning

Previous work related to pre-trained language models does not explicitly embed SMILES enumeration knowledge into language models during pre-training. Thereby, one of our proposed solutions is to encode that knowledge with a CL pre-training objective [Chopra et al., 2005]. Specifically, encoder-based BERT-like language models can be pre-trained to pull together representations of enumerations of SMILES to the latent representation of the canonical SMILES and vice versa.

Importantly, there is a limited body of work on pre-training SMILES-based language models with CL. Most of the prominent CL techniques use multiple molecular representations, train on GNNs, or combine them together [Yang et al., 2021b, Guo et al., 2022, Pinheiro et al., 2022, Wang et al., 2021].

In one of the prior works, vanilla transformer architecture which is essentially a seq2seq model has trained with CL objective [Shrivastava and Kell, 2021]. However, this work does not investigate BERT-like encoder-based models, which are currently state-of-the-art on QSAR and *Virtual Screening* tasks. Additionally, a more related work pre-trains the BERT language model with different pre-training tasks. The pre-training objectives include CL, atom, and molecule property prediction [Wu et al., 2022]. However, this work does not demonstrate the comparison of CL-based pre-training with currently more performant pre-training regimes, including MLM and physicochemical property prediction.

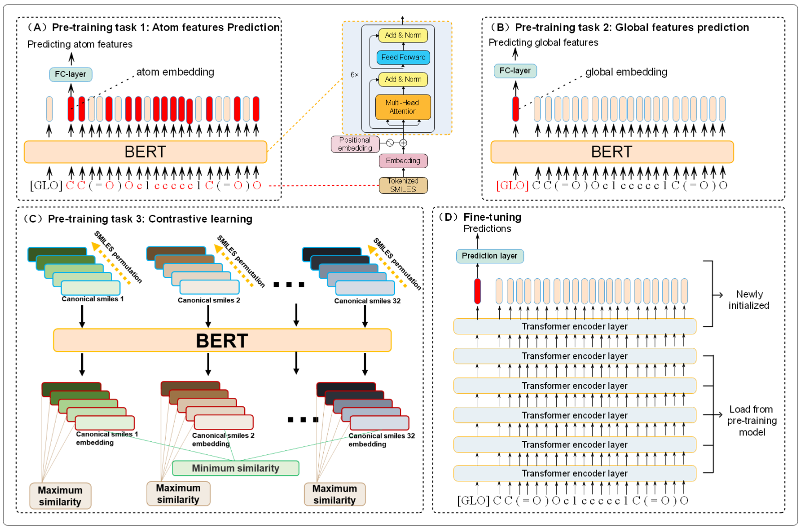
（A）Pre-training task 1: Atom features Prediction
Predicting atom features
FC-layer
atom embedding
BERT
[GLO] C C (= O) O c 1 c c c c c 1 C (= O) O
6×
Add & Norm
Feed Forward
Add & Norm
Multi-Head Attention
Positional embedding
Embedding
Tokenized SMILES
（B）Pre-training task 2: Global features prediction
Predicting global features
FC-layer
global embedding
BERT
[GLO] C C (= O) O c 1 c c c c c 1 C (= O) O
（C）Pre-training task 3: Contrastive learning
SMILES permutation
Canonical smiles 1
Canonical smiles 2
Canonical smiles 32
BERT
Canonical smiles 1 embedding
Canonical smiles 2 embedding
Canonical smiles 32 embedding
Minimum similarity
Maximum similarity
Maximum similarity
Maximum similarity
（D）Fine-tuning
Predictions
Prediction layer
Newly initialized
Transformer encoder layer
Load from pre-training model
[GLO] C C (= O) O c 1 c c c c c 1 C (= O) O

Figure 3.9: Training pipeline for Knowledge-based-BERT with contrastive learning for learning local and global fingerprints [Wu et al., 2022].

We model the task of learning SMILES representations analogous to learning sentence representations in NLP. Sentence representations can be learned in various pre-training settings. Exemplary techniques include predicting the surrounding sentences [Kiros et al., 2015], and training a siamese network on a *Natural Language Inference* (NLI) dataset [Cer et al., 2018, Yang et al., 2018].

Similarly, recent work has demonstrated that the sentence representations learned with MLM-based language models are anisotropic and are non-uniformly distributed [Ethayarajh, 2019a, Li et al., 2020a]. Thereby, the performance on downstream tasks such as document retrieval which is analogous to *Virtual Screening* and other related tasks suffer [Reimers and Gurevych, 2019].

Previous work to explicitly address the issues of alignment and uniformity with CL has demonstrated drastic improvements in learned sentence representations. One of the previous works proposes training a siamese BERT-like model with contrastive loss [Reimers and Gurevych, 2019]. The contrastive loss primarily maximizes the similarity between the fixed-length sentence representations. The fixed-length sentence vectors are obtained with one of the various strategies, including CLS-token embedding, or by performing a pooling operation.

Identically, another comparative approach involves feeding sentence pairs/triplets into the BERT-like transformer encoder [Gao et al., 2021]. The triplet input includes an anchor sequence, a positive sequence semantically identical to the anchor sequence, and a hard negative an unrelated sequence to the anchor. Then the encoder is trained by obtaining fixed sentence representations for each sentence similar to the earlier work [Reimers and Gurevych, 2019]. Thereafter, the similarity is maximized between the anchor and positive input and minimized between the anchor and the negative sentences with CL-based loss [Oord et al., 2018].

3.2 Low-data Molecular Drug Discovery

Deep neural networks are intrinsically data-hungry while learning generalizable patterns in data. Thereby, the availability of reasonably large datasets, especially for downstream supervised learning tasks, becomes critical. In this section, we discuss two different ways to tackle the low-data challenge for molecular discovery. First, we explain how SMILES enumerations can be seen as a viable data-augmentation mechanism. Subsequently, we also survey some of the applicable semi-supervised mechanisms that we later employ to deal with low-data molecular settings.

3.2.1 SMILES Enumerations as Data Augmentation

Data augmentation has been successfully applied to generate additional training instances [Taylor and Nitschke, 2018]. Additionally, data augmentation has also been shown as an implicit form of regularization to avoid overfitting. Importantly, data augmentation has proven to be pivotal in both computer vision [Shorten and Khoshgoftaar, 2019] and NLP domain [Shorten et al., 2021].

Similar to the other domains, molecular deep QSAR-based models require sizeable datasets to train. Enumerations can serve as a natural augmentation technique for molecular SMILES. This requires randomly scrambling the order of the atoms within the molecule [Bjerrum, 2017]. Hence, multiple SMILES variants can be produced from a single canonical SMILES. Such enumerations can be performed in a trivial manner with RDKit.

Including enumerations in training, data has proven to be advantageous for both discriminative and generative machine learning tasks with molecular SMILES [Chen and Tseng, 2021, Li et al., 2022].

Generative models trained with enumerated SMILES result in generating at least twice the amount of de novo molecules as the training data [Ertl et al., 2017, Arús-Pous et al., 2019]. Importantly, the generated SMILES have molecular properties as the training SMILES within that specific chemical space. Whereby, in another previous work it has been shown that enumerations-assisted generative models can achieve accuracy up to 98% for generating valid de novo molecules [van Deursen et al., 2020]. Lastly, generative models immensely benefit from enumerations in sparsely populated chemical spaces in low-resource scenarios [Skinnider et al., 2021].

Identically, enumerations are a useful resource for cheaply obtaining further labeled data for discriminative learning tasks in QSAR studies. Supervised discriminative learning models trained on enumerations-based augmented data have shown to significantly outperform models trained only on canonical SMILES [Tetko et al., 2019].

Lastly, it is essential to underpin the importance of data augmentation for low-data QSAR studies. Low data entails having very few training instances, constraining the training process. Previous studies have also demonstrated the quintessential characteristic of enumerations in such low-resource scenarios. SMILES enumerations as data augmentation are most effective when combined with transfer learning. Precisely, utilizing the augmentations in the SSL-oriented pre-training phase and then fine-tuning on the downstream low-resource dataset [Honda et al., 2019, Zhang et al., 2021b].

3.2.2 Semi-supervised Learning

Another important facet of our work is the application of transfer learning approaches to low-data regimes. Additionally, previous work pertaining to low-resource molecular property prediction with SMILES is scarce. Furthermore, some of the work such as [Honda et al., 2019] addresses this problem, however, the authors only employ a supervised learning approach which fails to generalize with very small labeled datasets. To alleviate such low-data regimes, we propose a fusion of transfer learning with semi-supervised learning for tackling low-data scenarios.

In our work, we employ prominent pseudo-labeling approaches from the semi-supervised learning paradigm [Yang et al., 2021a] appropriate for the molecular domain. We evaluate pseudo-label, which is a self-training method [Lee et al., 2013]. Pseudo label algorithm proposes training of a neural network with labeled and unlabeled data simultaneously. The overall neural network is trained in a supervised fashion. The unlabeled data is passed through the network and the maximum posterior from the predicted posteriors are considered pseudo labels. The training process of the pseudo label algorithm is shown in Fig. 3.10.

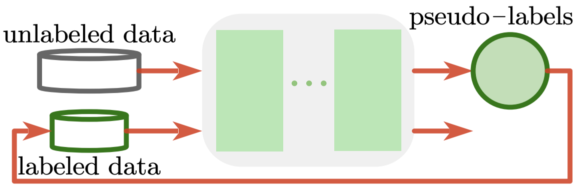

Figure 3.10: Pseudo label simultaneous training mechanism from [Yang et al., 2021a].

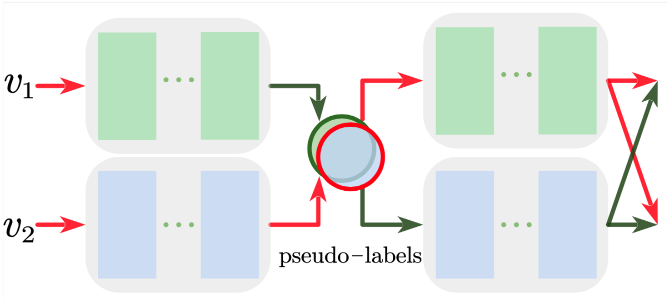

Figure 3.11: Illustration of the co-training process from [Yang et al., 2021a].

To complement the self-training approach, we also investigate the disagreement-based deep co-training approach [Blum and Mitchell, 1998]. Deep co-training essentially assumes that data has two complementary and different views. It proposes simultaneous training of two different networks for each of the two views, respectively. Finally, each view's network is used to label other view's unlabeled samples iteratively, until the unlabeled data is exhausted. An overview of the deep co-training pipeline is presented in Fig. 3.11.

3.3 Summary

The above-discussed approaches for molecular SMILES-based language modeling fail to effectively encode enumeration knowledge into the pre-trained language models. Hence, the learned molecular representations are susceptible to being low-fidelity. Importantly, the fidelity of the learned representations is directly correlated with the performance of downstream drug discovery models. Furthermore, training solely on the canonicalization algorithm's rules can be intrusive to the model's learning process.

In this thesis, we seek to alleviate the above shortcomings of the current SMILES language modeling mechanisms. First, we propose to incorporate enumeration knowledge for learning high-fidelity SMILES fingerprints. We employ the contrastive learning approaches proposed by [Gao et al., 2021, Reimers and Gurevych, 2019] to administer enumeration knowledge into learned latent fingerprints, which can potentially benefit the downstream QSAR Tasks. Whereby, the second part of our work is based on the coupling of semi-supervised learning approaches with pre-trained molecular language models to tackle low-data regimes for drug discovery, as described in Chapter 5.

Chapter 4
Datasets

Our proposed approaches use datasets that are utilized in the three stages of model training, namely pre-training, domain adaptation, and fine-tuning for downstream tasks. Importantly, the pre-training dataset is sampled from a general-purpose library of molecular SMILES. Whereby, domain adaptation requires an unlabeled dataset from the same domain as the downstream dataset. Finally, the model is fine-tuned on the downstream dataset in a semi-supervised/supervised learning manner.

This chapter shows the datasets used in the model training process at the three previously discussed stages and their different statistics.

4.1 Datasets Gathering

This section explains the datasets with respect to the corresponding tasks, these datasets are used for the training and evaluation of our proposed molecular language models.

4.1.1 Pre-training

We pre-train the proposed language models on the GuacaMol [Brown et al., 2019b] benchmark dataset. The GuacaMol dataset is a subset of ∼1.6M molecules sampled from the ChEMBL database [Gaulton et al., 2017]. This dataset primarily serves in learning domain-agnostic molecular representations with different pre-training objectives. Also, the detailed statistics about the GuacaMol dataset are presented in Tab. 4.1.

4.1.2 Virtual Screening

We use the virtual screening benchmarking platform by RDKit [Riniker and Landrum, 2013] for *Virtual Screening* evaluation. The benchmarking platform evaluates the molecular fidelity of fingerprints on 69 datasets. Moreover, each dataset is composed of a single drug-protein target. We evaluate representations learned by different pre-training objectives directly on *Virtual Screening* using this benchmark. Specifically, for a fixed number of query SMILES (number of query SMILES = 5), we evaluate how accurately can the representations from language models retrieve the ground truth of active molecules.

4.1.3 QSAR

We use a subset of datasets related to the Biophysics and Physiology domain from the MoleculeNet benchmark [Wu et al., 2018] for evaluating the pre-trained language models on downstream QSAR tasks. Importantly, all the datasets in the MoleculeNet benchmark employ scaffold splitting, which is effective in practical applications. Below, we outline the specific details about each of the QSAR datasets.

MUV

The *Maximum Unbiased Validation* (MUV) dataset is used for benchmarking virtual screening methods. It consists of compounds with 17 tasks filtered from the PubChem bioactivity data using the refined nearest neighbor analysis [Rohrer and Baumann, 2009]. In our work, we employ the MUV dataset for the intermediate domain adaptation prior to fine-tuning and evaluating QSAR and the *Virtual Screening* tasks.

BACE

The BACE dataset consists of a collection of *human β-secretase 1* (BACE-1) inhibitors with their corresponding IC50 and binary binding results [Subramanian et al., 2016]. It is essentially compiled by aggregating the experimental values from the existing studies. The MoleculeNet benchmark contains the BACE compounds with their 2D structures and binary labels.

BBBP

The *Blood–brain Barrier Penetration* (BBBP) dataset is based on the modeling of the permeability brain barrier of drugs, hormones, and neurotransmitters [Wu et al., 2018]. Certain nervous system issues are caused once this barrier is breached by the nervous system drugs. This dataset contains a library of compounds with their brain permeability properties captured in binary labels.

Tox21

The Tox21 dataset was created as the result of the *Toxicology in the 21st Century* (Tox21) initiative for collecting toxicity measurements. It contains a collection of 12 tasks for qualitative toxicity measurements.

ClinTox

The ClinTox dataset was created during the curation of the MoleculeNet benchmark [Wu et al., 2018]. It specifically compares the toxicity measurements of *Food and Drug Administration* (FDA) approved drugs against the drugs which failed their clinical trials. Hence, the dataset includes the toxicity measurements of compounds and their FDA approval status as two separate labels.

4.1.4 Datasets Summary

The following Tab. 4.1 contains the summary of the datasets alongside their evaluation metrics. Furthermore, detailed information about the evaluation is provided in Sec. 6.1.3.

Categeory	Dataset	Tasks	Type	Mols	Metric
–	GuacaMol	–	P	1,273,104	–
Biophysics	MUV	–	D	93,127	–
	BACE	1	C	1,522	AUROC
Physiology	BBBP	1	C	2,053	AUROC
	Tox21	12	C	7,831	AUROC
	ClinTox	2	C	1,491	AUROC

Table 4.1: Summarized information of MoleculeNet and GuacaMol [Brown et al., 2019b]. “C", “P", and “D" in the type column indicates classification, pre-training, and domain adaptation (for details about domain adaptation refer to Sec. 5.1.1) respectively.

4.1.5 Application: HIPS Dataset

In addition to the public MoleculeNet benchmark, we also evaluate our proposed approaches on a real-world propriety dataset by *Helmholtz Institute for Pharmaceutical Research Saarland* (HIPS). Importantly, the labeled dataset comprises molecule cytotoxicity measurements. To determine cytotoxicity two samples of a cell line are compared where one is treated and the other one is untreated over measured relative growth inhibition [Webel et al., 2020]. We consider a molecule to be cytotoxic if the growth of the treated sample is inhibited by at least 50% as compared to the untreated one. Additionally, in our work, we use the cell line **HepG2** at the concentration level of 100 µM.

The overall dataset consists of 6673 molecules including the labeled ones. Whereby, 248 of the molecules are labeled. We perform a random split in this case of 64% (158 molecules), 16% (40 molecules), and 20% (50 molecules) of the train, validation, and test dataset respectively.

4.2 Dataset Splitting

In a standard machine learning setting, a dataset is typically randomly split into train/validation/test datasets. Similarly, the MoleculeNet datasets are split into these three splits with proportions 80/10/10 respectively. However, in the context of molecular datasets random splits are not always ideal for the machine learning algorithm [Wu et al., 2018].

The MoleculeNet datasets are provided with various splitting mechanisms. In our work, we use scaffold splitting, which entails splitting based on the 2D structural forms as provided by the RDKit. The choice of scaffold splitting mechanism was made as it provides more practically useful information about the protein target than random splitting. A visual illustration highlighting the difference between the random and scaffold split is presented in Fig. 4.1.

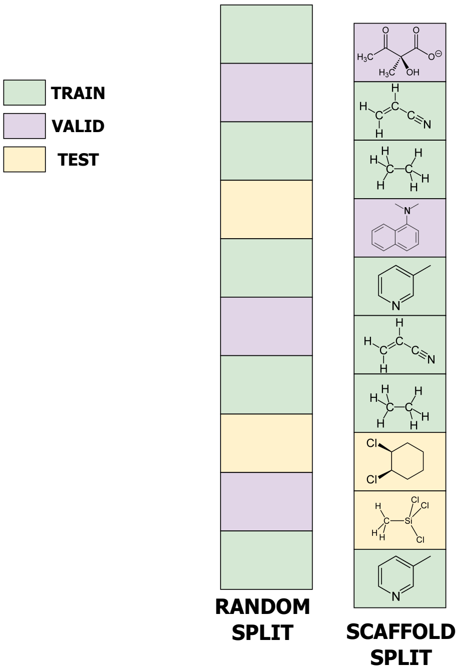

Figure 4.1: Dataset Splitting

Chapter 5
Methodology

This chapter describes the intricate details of our proposed approaches for training molecular language models robust to low-data settings, leveraging high-fidelity learned representations. Hence, this chapter comprises of two parts. The first part explains our proposed extensions to existing molecular language model pre-training regimes for representation learning. Whereby, the second half elicits the adaption of the semi-supervised learning paradigm to alleviate low-data scenarios.

Pre-training encompasses the representation learning facet of the neural language models. The current SMILES-based language models use the NLP-inspired MLM pre-training [Chithrananda et al., 2020]. Additionally, recent state-of-the-art approaches have adapted pre-training using the physicochemical properties in an MTR fashion [Fabian et al., 2020]. We select the base model, which is pre-trained with MLM. In Sec. 5.1, we propose two alternative pre-training approaches to incorporate enumeration knowledge for learning rich representations.

In Sec. 5.1.1, we describe our proposed novel domain adaptation framework for BERT-like language models. The domain adaptation entails multi-task pre-training, simultaneously combining MLM, MTR, and CL objectives. Alternatively, in Sec. 5.1.2, we propose a denoising-based encoder-decoder language model for SMILES canonicalization. Here, the language model learns to canonicalize the noisy input SMILES.

The transformers-based language models are inherently data-hungry. Thereby, their performance is hindered significantly in low-data settings. In Sec. 5.2, we discuss mechanisms to alleviate such scenarios. Concretely, we assume SMILES enumeration is a quintessential data augmentation mechanism. Furthermore, we leverage additional unlabelled training data with semi-supervised learning to enhance model generalization.

5.1 SMILES Representation Learning

As discussed in Sec. 2.1, SMILES is a simple language for molecules. Thereby, enumeration is an essential property of SMILES representation. However, the existing SMILES language modeling mechanisms fail to explicitly encode enumeration awareness. Thus, the absence of a mechanism that incorporates enumeration knowledge during pre-training leads to performance degradation on downstream drug discovery tasks such as QSAR [Payne et al., 2020].

We propose to incorporate enumeration knowledge into language models with two alternative methods for pre-training language models. We alter pre-training for both transformer-based encoder and encoder-decoder architectures. For BERT-like language models, we propose a three-step domain adaptation framework, contrary to two-step transfer learning approaches. In Sec. 5.1.1, we introduce an additional intermediate domain-specific pre-training called domain adaptation. During domain adaptation, we incorporate enumeration knowledge into a pre-trained language model via CL.

Conversely in Sec. 5.1.2, we elaborate that the purpose of our proposed encoder-decoder model pre-training is to learn a neural canonicalization function using denoising. Specifically, the input SMILES-based molecules are corrupted with noising functions including enumeration, and token masking. The decoder is then optimized to output the canonical version of the noisy SMILES inputs. Hence, the resultant pre-trained language model possesses inherent implicit knowledge about enumerated SMILES. Thereby, it maximizes the likelihood of producing similar latent representations for the same molecule when it is enumerated.

5.1.1 Contrastive Encoder Transformer

We extend the standard transfer learning regime with our proposed BERT-based molecular domain adaptation framework, which consists of the following three stages:

- **Pre-training:** Our proposed molecular domain adaptive framework preserves the existing pre-training mechanisms of MLM and MTR. Importantly, we only use either MLM or MTR at a time to pre-train a language model.
- **Domain Adaptation:** At the second stage, we propose additional intermediate pre-training on domain-specific unlabeled SMILES with multi-task self-supervised objectives including MLM, MTR, and CL.

- **Fine-tuning:** Finally, the domain-adapted language model is fine-tuned in a supervised fashion on the downstream QSAR task and then evaluated. Whereby, for the *Virtual Screening* domain-specific latent representations of SMILES are directly used without any additional training.

Step 1: Pre-training

Identical to the standard SMILES-based BERT pre-training, we two employ MLM and MTR SSL pre-training objectives, as shown in Fig. 5.1. In MLM, 15% of the input SMILES tokens are randomly masked. The language model then outputs the probability distribution over all possible tokens in the vocabulary for each masked token. This process enables the language model to learn the semantic structure and produce fingerprints that encode the relationship between SMILES tokens conditioned on the context.

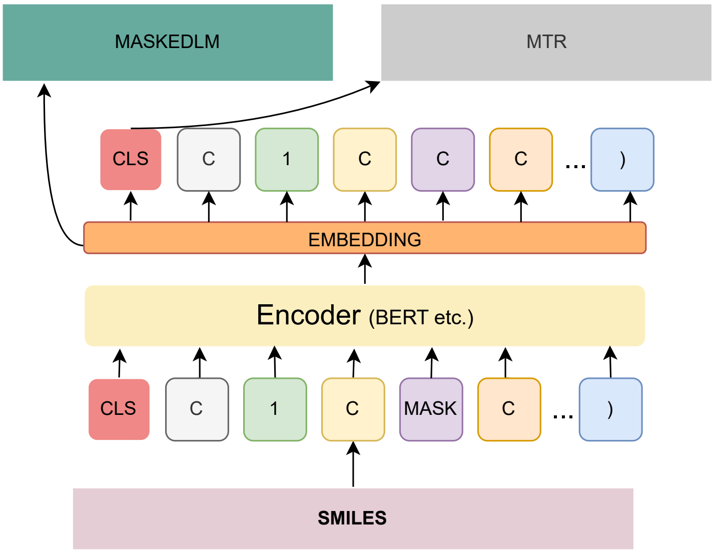

Figure 5.1: (Step 1) Pre-training on a large unlabelled dataset.

MLM pre-training pertains to minimizing the cross-entropy loss, as specified in Eq. (5.1). Specifically, the loss function compares the difference between the predicted probability distribution with the ground truth tokens distribution. Furthermore, the language model parameters are then updated using an optimization algorithm such as *Stochastic Gradient Descent* (SGD) [Amari, 1993].

$$L_{MLM} = -\sum_{c=1}^{M} y_{o,c} \log(p_{o,c}) \tag{5.1}$$

Conversely, the MTR pre-training is independent of the linguistic structure of the SMILES.

The MTR-based pre-training involves the simultaneous prediction of the 220 real-valued physicochemical properties of the input molecule. Hence, the language model learns implicit chemical characteristics associated with a given molecule. In our work, we use the RDKit framework to acquire the physicochemical descriptors of the pre-training dataset. These descriptors are part of the following categories:

- FRAGMENT
- CHARGE
- GENERAL
- SURFACE
- GRAPH
- SIMPLE
- DRUGLIKENESS
- ESTATE
- LOGP
- REFRACTIVITY
- GENERAL

The BERT-like language model uses multi-task mean squared error loss as specified in Eq. (5.2), where D is the 220-dimensional chemical descriptors and N is the number of training samples. For smooth convergence, we use mean and standard deviation to normalize each descriptor as a pre-processing step.

$$L_{MTR} = \sum_{i=1}^{N} \sum_{j=1}^{D} (o_{ij} - y_{ij})^2 \tag{5.2}$$

Step 2: Domain Adaptation

The domain adaptation stage enhances the pre-trained latent SMILES representations twofold. First, by injecting domain-specific knowledge about the downstream chemical space. Second, we introduce the CL objectives as shown in Fig. 5.2, to infuse enumeration awareness into the learned representations. Thereby, the domain adaptation step re-uses pre-training objectives of MLM and MTR for learning domain-specific knowledge. Importantly, incorporating CL-based enumeration awareness in the second step makes our approach more effective and data-efficient. As we will demonstrate empirically in Chapter 6, the size of the domain-specific dataset (*MUV*) is significantly smaller than the pre-training (*GuacaMol*) dataset.

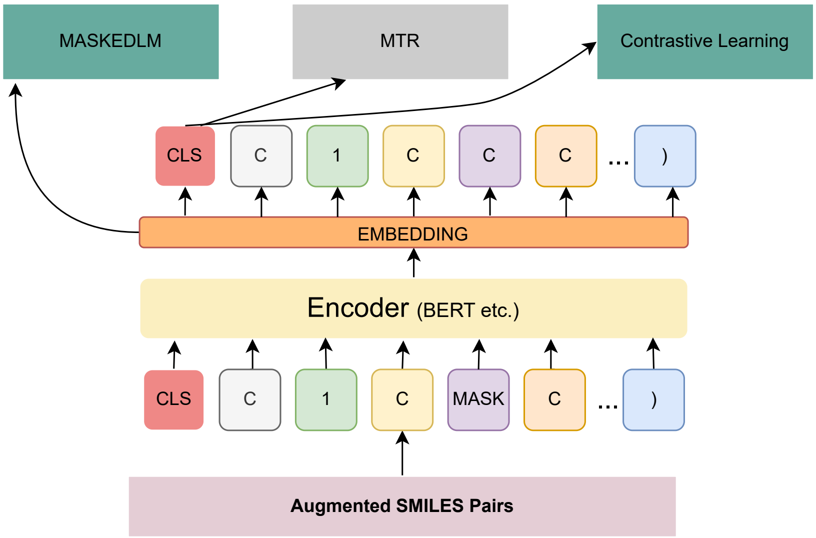

Figure 5.2: (Step 2) Domain adaptation with downstream domain dataset.

We employ two different architectures to incorporate the CL objectives as part of the domain adaptation step, namely *Siamese BERT* (SBERT) and *Contrastive BERT* (CBERT). The SBERT architecture [Reimers and Gurevych, 2019] uses two BERT-like encoders. The encoders are presented with input SMILES pairs. A SMILES pair can have two possible combinations. It can have a canonical SMILES and one of its enumerations. Alternatively, a canonical and a random SMILES from the dataset, also known as hard negative, can form the pair.

The encoders learn the latent representations of SMILES pair and then minimize or maximize based on the relation between the inputs. As shown in Fig. 5.3, one of the encoders exclusively learns latent representations for the canonical SMILES. Whereby, the other encoder translates the enumerations and hard negatives to the latent space. We use the multiple negative ranking loss function [Henderson et al., 2017] to train the SBERT.

The multiple negatives ranking loss function performs well to learn representations of a canonical SMILES and its enumerations to be in a closed vicinity within the chemical latent space. We use fixed-length embeddings r_c and r_e taken from the [CLS] token prediction as latent representations of the canonical SMILES and its enumeration. The loss function enforces weight updates of the language model such that a canonical SMILES and its enumeration are pulled closer in the underlying chemical latent space and vice versa for hard negatives. The mathematical formulation of the loss function is shown in Eq. (5.3).

$$L(r_c, r_e) = -r_c^T r_e + \log \sum_{n \in S_c} \exp(r_c^T r_n) \tag{5.3}$$

- r_c: Latent representation of the canonical SMILES.
- r_e: Latent representation of the enumerated SMILES.
- r_n: Latent representation of the hard negative SMILES.
- S_c: Set of randomly sampled canonical SMILES to create hard negative pairs.

Precisely, this is achieved by changes to the model parameters θ_c and θ_e that maximizes the similarity between r_c and r_e and conversely, minimize the similarity between r_c and r_n. The loss function used to achieve this is shown in Eq. (5.3), where S_c is a set of in-batch randomly sampled SMILES used as hard negatives.

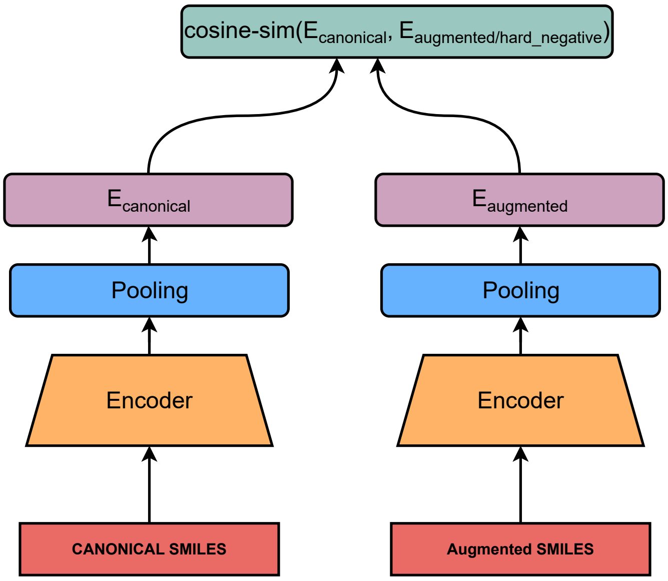

Figure 5.3: Siamese BERT architecture with contrastive learning.

The CBERT model [Gao et al., 2021] takes SMILES triples instead of pairs. Each triple consists of canonical SMILES, enumeration, and a hard negative. As shown in Fig. 5.4, the encoder is based on a single BERT model. Subsequently, the latent SMILES representations of the canonical and enumerated SMILES are pulled together, whereby, the latent representation of the hard negative is pushed away from the latent representations of the other two SMILES simultaneously. Hence, CBERT only requires a single forward pass of a BERT-like model to perform both maximization and minimization between the SMILES representation based on their nature. Unlike the SBERT where hard negatives are sampled within the batch during the forward pass on the fly, the hard negatives in a

SMILES triple are fixed and are sampled prior to training the CBERT.

The CBERT model uses the contrastive framework [Chen et al., 2020] which combines the cross entropy with in-batch negatives [Chen et al., 2017]. It assumes that a dataset with a set of tuples of text phrases $\mathcal{D} = \{(x_i, x_i^+)\}_{i=1}^N$ where each pair has similar connotations. We transfer this analogy to the molecular SMILES dataset, where x_i refers to the canonical SMILES and x_i^+ represents the enumeration of x_i. Formally, the contrastive loss for the CBERT is expressed as follows:

$$L(r_c, r_e, r_n) = -\log \frac{e^{sim(r_c, r_e)/\tau}}{\sum_{j=1}^{M} e^{sim(r_c, r_n)/\tau'}} \tag{5.4}$$

- $sim(r_1, r_2)$: The cosine similarity function defined as $\frac{r_1^T r_2}{\|r_1\| \|r_2\|}$.
- r_c: Latent representation of the canonical SMILES.
- r_e: Latent representation of the enumerated SMILES.
- r_n: Latent representation of the hard negative SMILES.
- M: Number of SMILES triples in the mini-batch.
- τ: A temperature hyper-parameter that is used to control the entropy.

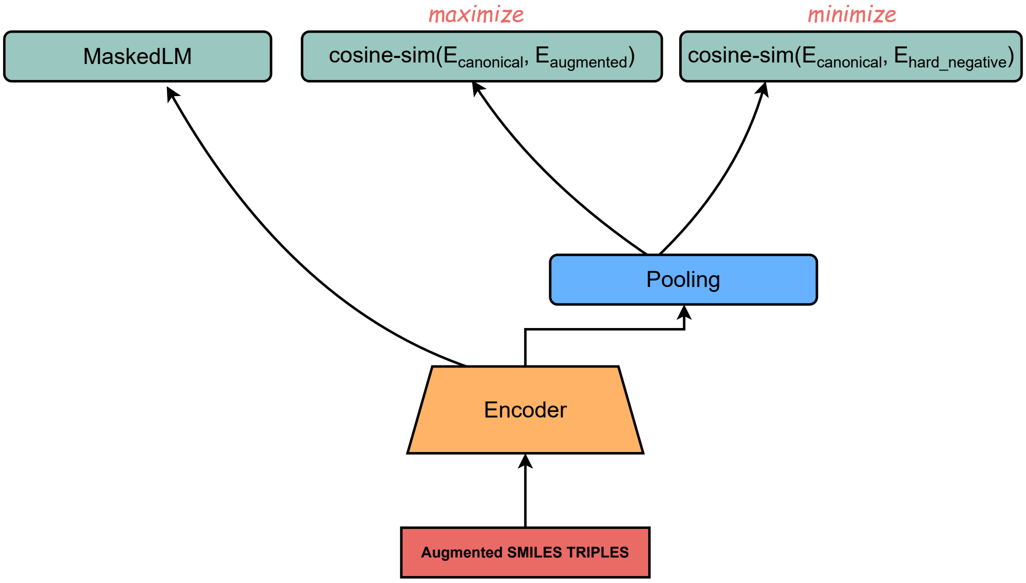
maximize
minimize
MaskedLM
cosine-sim(E_canonical, E_augmented)
cosine-sim(E_canonical, E_hard_negative)
Pooling
Encoder
Augmented SMILES TRIPLES

Figure 5.4: Contrastive BERT architecture with SMILES triples.

Furthermore, the contrastive framework infuses two implicit properties into the learned latent fingerprints, (1) distribution, and (2) alignment. The distribution property states that learned representation should be uniformly distributed. However, representations learned with the recent deep language modeling approaches suffer from the anisotropy problem [Ethayarajh, 2019b, Li et al., 2020b]. The anisotropy problem states that the learned dense representations follow a cone shape in the latent space. However, the contrastive learning objective appears to alleviate the anisotropy problem [Gao et al., 2021]. Moreover, the alignment property entails that the positive SMILES pairs should be closer to each other distance-wise compared to the hard negatives. We also empirically show this in Chapter 6, that the CL-based objectives indeed improve the alignment of learned fingerprints.

In summary, we introduce a three-faceted multi-task self-supervised domain adaptation phase which includes (1) MLM, (2) MTR, and (3) *Contrastive Learning*. The domain adaptation allows the language model to produce domain-specific molecular fingerprints relevant to the underlying chemical space. Furthermore, CL-based training infuses enumeration knowledge into a pre-trained language model in a data-efficient manner. Critically, the success of the domain adaptation heavily relies on the choice of dataset. We evaluate our approaches on both a proxy dataset and the downstream dataset and discuss the implications of selecting either of them in Chapter 6.

Step 3: Fine-tuning

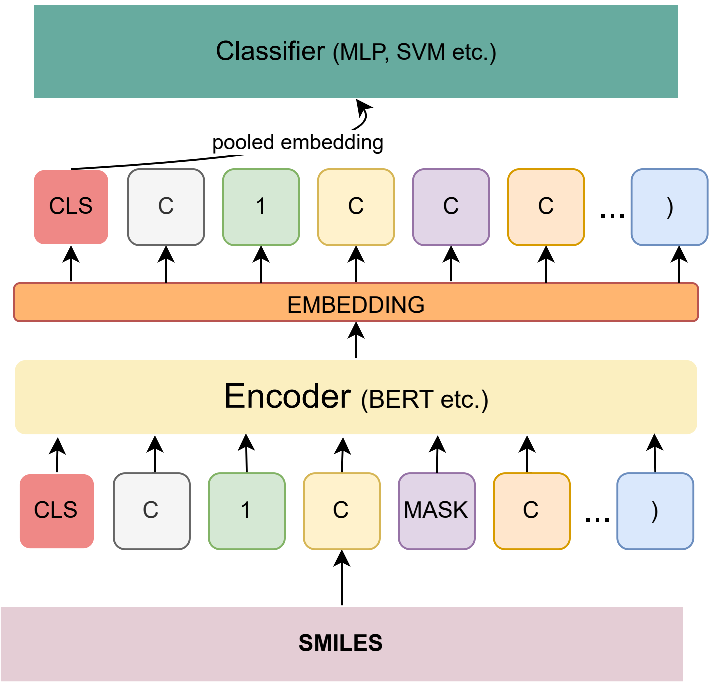

Figure 5.5: (Step 3) Supervised fine-tuning of the domain-adapted language on the downstream dataset.

We perform the standard supervised fine-tuning at the final step of our proposed molecular domain adaptation framework, as shown in Fig. 5.5. Hence, this step requires completely labeled datasets. In our work, we freeze the weights of the BERT encoder and extract dense fingerprints using the encoder. Here, freezing refers to we no longer update the weights of the model. Subsequently, we train multi-label *Support Vector Machine* SVM [Cortes and Vapnik, 1995] models for each dataset with the extracted fingerprints as input features. Lastly, we evaluate our trained classifiers as described in Chapter 6.

5.1.2 Denoising Encoder-decoder Transformer

In addition to encoder-based transformer architectures, we also investigate a transformer-based encoder-decoder architecture. Traditionally, encoder-decoder SMILES language models are pre-trained as autoencoders, with SMILES reconstruction as the pre-training objective. Thereby, such models can produce representations of enumerated SMILES of a molecule that are distant in the molecular latent space.

To address this for encoder-decoder models, we propose to replace the vanilla SMILES reconstruction objective with a denoising-based reconstruction objective. Primarily, we build our work on the BART transformer architecture [Lewis et al., 2019]. The BART model was originally trained on a natural language corpus. The pre-training phase is based on corrupting the input documents with a finite set of stochastic noising schemes, as described in Sec. 2.3.2.

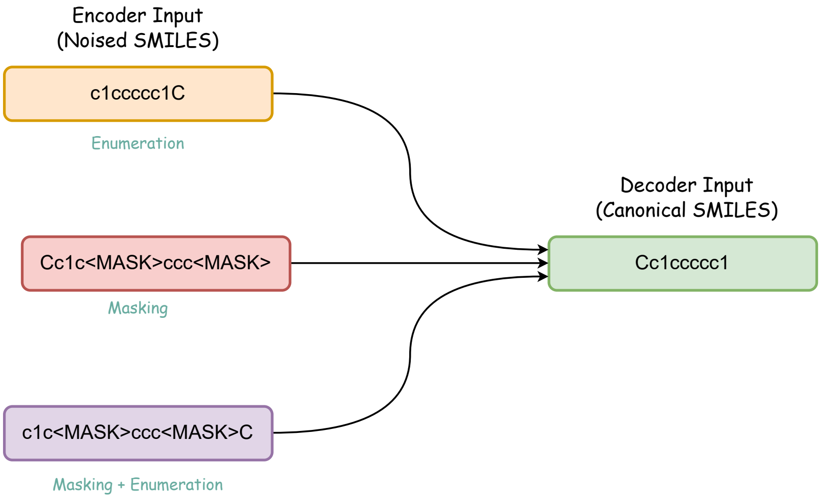

Figure 5.6: SMILES Noising Schemes

Most of the noising schemes prescribed in the original work are based on natural language. Hence, the original noising functions are inapplicable to SMILES representation.

Consequently, we introduce enumeration as a noising scheme, where input SMILES is enumerated, and the model learns to canonicalize the input SMILES. Additional noising schemes include masking tokens in the input SMILES and combining both enumeration and masking schemes. A summarized illustration of the encoder and decoder input (noising functions and output) is presented in Fig. 5.6. Finally, the described noises are induced into the canonical SMILES as part of the data preparation step. Then the model is trained to predict the canonical SMILES whilst training with a denoising-based reconstruction objective.

The BART model follows the architecture of the vanilla transformer sequence-to-sequence architecture [Vaswani et al., 2017]. The only difference is the choice of the activation, where BART uses GeLUs [Hendrycks and Gimpel, 2016] instead of the ReLU activation functions. The architecture is composed of a bidirectional encoder and an autoregressive decoder, as shown in Fig. 5.7. The BART model is trained by optimizing the reconstruction loss, which is the cross-entropy loss between the output of the decoder and the original noise-free canonical SMILES. Furthermore, the overall architecture is similar to the BERT architecture except for the following differences:

- No feed-forward layer prior to word prediction.
- Cross-attention is applied over each layer of the decoder, with the output of the final hidden layer of the encoder.

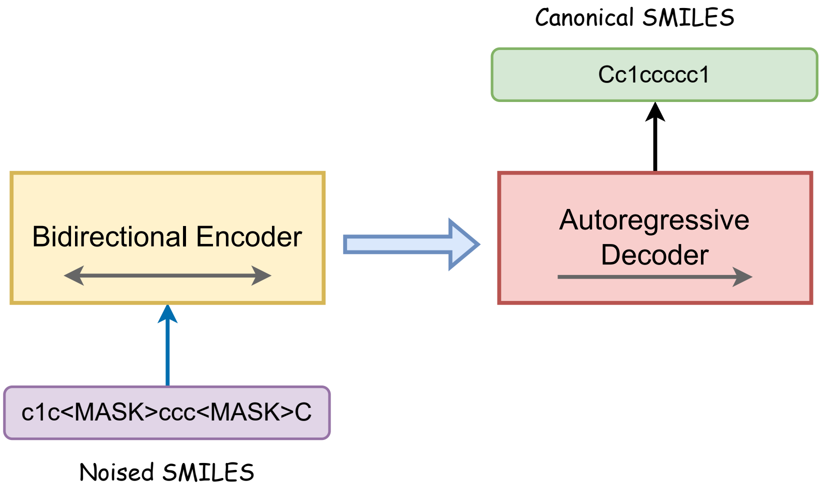

Figure 5.7: BART Transformer Architecture

Succinctly put, we present an alternative approach for infusing enumeration knowledge into a pre-trained sequence-to-sequence language model. Unlike, the encoder-based domain adaptation framework, the BART-based model incorporates enumeration during pre-training. Hence, the model can directly be fine-tuned immediately after pre-training. For fine-tuning the decoder last hidden output is used as a SMILES fingerprint for training the downstream SVM-based classifiers.

5.2 Molecular Low-Data Regime

The transformer-based pre-trained language models have demonstrated state-of-the-art performance on various molecular property prediction QSAR tasks [Ahmad et al., 2022, Fabian et al., 2020]. However, molecular dataset preparation requires rigorous vetting and validation from domain experts. Hence, this exercise levies a significant amount of human and computational resources. Furthermore, accumulating a validated library of molecules adhering to specific properties or affinities to molecules adds additional non-trivial constraints. Consequently, this hinders the application of supervised fine-tuning of pre-trained language models.

5.2.1 Enumeration-aware Semi-supervised Learning

We propose adopting the well-established paradigm of *Semi-supervised Learning* (SESL) to circumvent the low-data challenges. Using the SESL-based approaches implies we can leverage both low-quantity labeled and large unlabeled datasets simultaneously (see Fig. 5.8). Specifically, we re-purpose two prominent pseudo-labeling-based SESL approaches, (1) Pseudo-label [Lee et al., 2013] and Deep Co-training [Blum and Mitchell, 1998]. These pseudo-labeling techniques use predictions of the downstream classifier to label the unlabeled samples, known as pseudo-labels. After that, the classifier's weights are updated using both labeled and unlabeled samples in a supervised fashion.

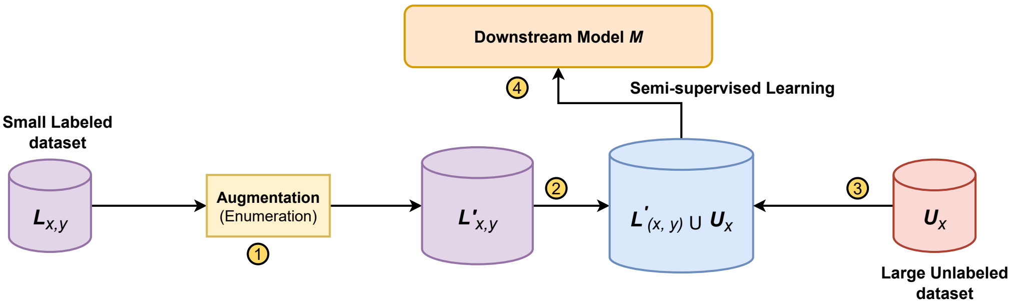

Figure 5.8: Semi-supervised learning pipeline for low-data molecular settings.

It is important to note this application of the SESL only modifies the fine-tuning phase of the pre-trained language model. Whereby, the rest of the transfer learning pipeline for both enumeration-aware encoder-only (see Sec. 5.1.1) and encoder-decoder (see Sec. 5.1.2) architectures remains unchanged. Resultantly, these SESL approaches are able to benefit from enumeration-aware fingerprints in low-data regimes.

Pseudo-label

Pseudo-label is a simple data-efficient version of the semi-supervised algorithm for training deep neural networks. The algorithm simultaneously uses both supervised samples to train the network and then uses the high-confidence predictions on the unlabeled samples as pseudo labels [Lee et al., 2013], as shown in Fig. 5.9. Subsequently, the model is trained in a supervised manner on both labeled and (pseudo-labeled) originally unlabeled examples, with one caveat. The pseudo-label loss function as defined in Eq. (5.5) controls the proportion of the contribution of the unlabeled examples to overall loss. This is achieved by introducing a loss weighting function $\alpha(t)$.

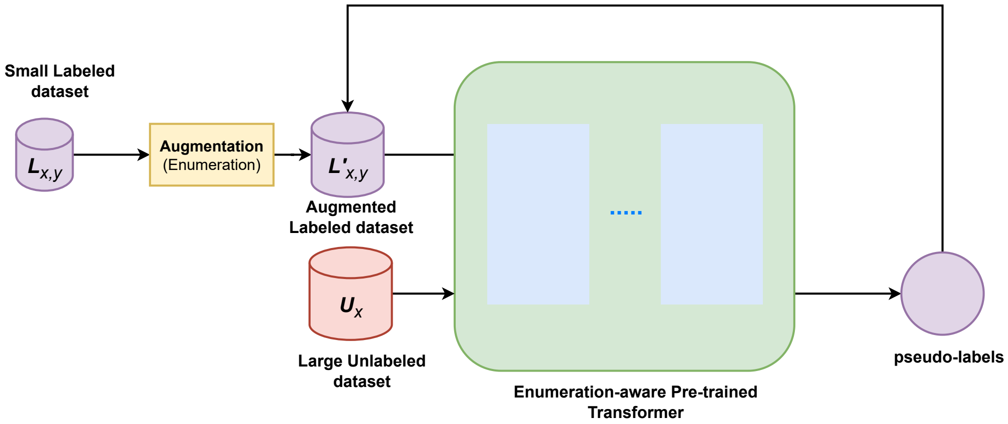
Small Labeled dataset
$L_{x,y}$
Augmentation (Enumeration)
$L'_{x,y}$
Augmented Labeled dataset
U_x
Large Unlabeled dataset
.....
Enumeration-aware Pre-trained Transformer
pseudo-labels

Figure 5.9: Training pipeline of the Pseudo-label SESL approach.

$$L(X,y) = \frac{1}{n}\sum_{m=1}^{n}\sum_{i=1}^{C}\mathcal{R}(y_i^m, f(x_i^m)) + \alpha(t)\frac{1}{n'}\sum_{m=1}^{n'}\sum_{i=1}^{C}\mathcal{R}(y_i^{'m}, f(x_i^{'m})) \tag{5.5}$$

- C: Total number of classes in the dataset.
- m, m': Number of samples in the labeled and unlabeled batches respectively.
- $f(x)$: The pre-trained molecular language model.
- $\mathcal{R}(y, f(x))$: The standard cross-entropy loss function over ground-truth and predictions.
- $\alpha(t)$: The alpha weight function is used to control the proportion of change unlabeled samples make to the overall loss. It increases as the epochs increase allowing unlabeled samples to make larger contributions to loss during later epochs. We can mathematically describe the $\alpha(t)$ function as following:

$$\alpha(t) = \left\{ \begin{array}{ll} 0, & \text{if } t < T_1 \\ \frac{t-T_1}{T_2-T_1}\alpha_f, & \text{if } T_1 \le t < T_2 \\ \alpha_f, & \text{if } T_2 \le t \end{array} \right\} \tag{5.6}$$

Here, we use the following values of the parameters of the alpha function, which include $\alpha_f = 3$, $T_1 = 20$, and $T_2 = 60$. Whereas, t argument to the $\alpha(t)$ function refers to the current value of epoch.

Deep Co-training

The deep co-training framework is based on the key assumption that every example in the dataset has two distinct views, which are also complementary. This entails each view is sufficient to train a performant classifier independently. Hence, deep co-training trains two different classifiers, one for each view respectively, as shown in Fig. 5.10. The two classifiers are used to acquire pseudo-labels, which are predictions with high posteriors. These pseudo-labels are subsumed into the labeled dataset. This process is iteratively repeated until the unlabeled dataset is exhausted or a pre-defined stopping criterion is met [Blum and Mitchell, 1998].

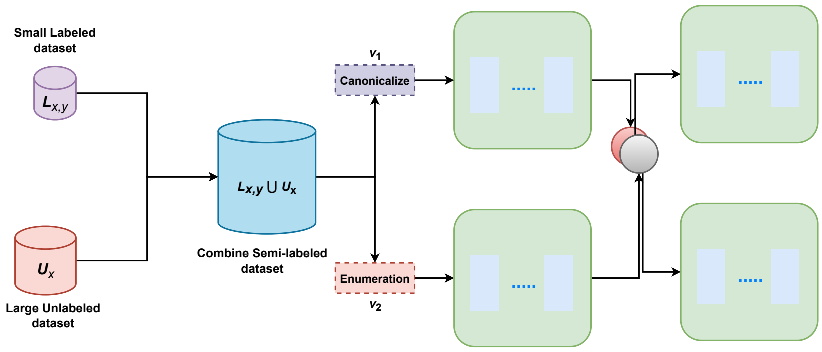

Figure 5.10: Deep Co-training pipeline for training with a semi-labeled dataset.

We adapt co-training for dealing with low-data regimes by defining their mapping in terms of two complementary views of SMILES. We define the canonical SMILES as *view-1* (v_1). Whereby, the enumeration of the same canonical SMILES is seen as *view-2* (v_2). Additionally, the deep co-training loss consists of two sub-losses influenced by labeled and unlabeled examples respectively. Furthermore, the overall loss of the two classifiers C_1, and C_2 is jointly optimized. The overall loss for the deep co-training is defined as follows:

$$L(v_1, v_2, y) = L_{sup} + \lambda L_{cot} \tag{5.7}$$

- L_{sup}: Corresponds to the standard supervised cross-entropy loss for each classifier respectively.

$$L_{sup} = \mathcal{R}(y, C_1(v_1)) + \mathcal{R}(y, C_2(v_2)) \tag{5.8}$$

- λ: It is a hyperparameter used to control the contribution of the co-training loss of unlabeled examples.

- L_{cot}: This term ensures that both classifiers agree on their predictions. Concretely, the predictions of both classifiers on each view for unlabeled data should be as similar as possible. Therefore, this can be measured using the Jensen-Shannon divergence as expressed below:

$$L_{cot} = \mathcal{R}(\frac{1}{2}(C_1(v_1)) + C_2(v_2))) - \frac{1}{2}\mathcal{R}(\frac{1}{2}(C_1(v_1)) + C_2(v_2))) \tag{5.9}$$

5.3 Summary

In this chapter, we proposed to extend current transfer learning approaches into two dimensions.

(1) Incorporating the enumeration property of molecular SMILES into learned latent fingerprints.

(2) Introducing SESL as a viable solution for low-data regimes. Specifically, alleviating QSAR settings with non-trivial requirements for a labeled dataset with pre-defined properties or affinities to molecules.

To achieve enumeration awareness, we propose to alter the pre-training phase of both encoder-only and encoder-decoder transformer-based language models. For the BERT-like encoder-only transformer, we introduced contrastive learning as an intermediate self-supervised pre-training objective. Whereby, for encoder-decoder-based BART architecture, we presented novel noising functions for denoising-oriented optimization. Whereby, for low-data settings, we proposed the adoption of pseudo-labeling-based variants of SESL. Specifically, we explained how pseudo-label, and deep co-training can be applied to a SMILES-based low-data molecular dataset. In Chapter 6, we explain the experimental settings, evaluation metrics, and results of our proposed methods. Importantly, we discuss the experimental results in light of the research question presented in Sec. 1.4 and the noteworthy takeaways.

Chapter 6
Experimental Evaluation

This chapter explains the experimental outcomes accompanied by the experimental setup of our proposed approaches. Furthermore, the described experiments directly address the research questions presented in Sec. 1.4. We also perform additional analysis discussing the evaluation results. The primary objectives of the experiments are twofold, (1) evaluating enumeration-aware language models, (2) adapting enumeration-aware language models to low-data settings.

The evaluation of enumeration-aware language models entails whether incorporating enumeration knowledge enhances their performance on the downstream tasks. Here, the downstream tasks include QSAR and *Virtual Screening*. We also assess explicit enumeration awareness in the latent space of the learned molecular fingerprints by setting up an additional independent experiment. Finally, we investigate the possible ways of adapting the enumeration-aware language models in low-data settings. Specifically, we examine the introduction of SESL approaches to mitigate the scarcity of labeled data points.

6.1 Experimental Setup

In this section, we explain the baseline models, model configurations, and evaluation metrics employed in our experiments.

6.1.1 Baselines

In our work, we use the baseline self-supervised model pre-trained with the MLM objective on molecular SMILES [Ahmad et al., 2022, Fabian et al., 2020]. This approach entails pre-training a transformer model by randomly masking a small fraction of the input. Here, the training involves minimizing the cross-entropy loss for all predicted masked tokens. Further details about this approach are specified in Sec. 5.1.

For the evaluation of downstream tasks, the process varies with respect to the nature of the task. For instance, for QSAR tasks, we extract the SMILES representations as fingerprints from the pre-trained language models using the downstream labeled dataset. Subsequently, identical to the work of [Winter et al., 2019], we train an additional *Support Vector Machine* (SVM) classifier with the extracted fingerprints as input features. However, for the *Virtual Screening* task, we solely extract the fingerprints and then directly evaluate using the virtual screening benchmarking platform by RDKit [Riniker and Landrum, 2013].

6.1.2 Model Configurations

We evaluate two different architectures namely CBERT and SBERT to inject enumeration in learned fingerprints as introduced in Sec. 5.1. Hence, we used a different set of hyperparameters for each architecture, respectively. The necessary details about the hyperparameters are provided in Tab. 6.1.

Hyperparameters	**CBERT**	**SBERT**	**BART**
Batch Size	32	16	32
Learning Rate	0.00005	0.00005	0.0001
Number of Epochs	10	10	20
Maximum Input Length	32	32	32
Pooling Strategy	CLS	Mean	-

Table 6.1: Summary of hyperparameters for SBERT and CBERT architectures.

Notably, the pooling strategy corresponds to the mechanism used to extract latent representations prior to applying the contrastive learning objectives. For CBERT, the CLS token's hidden representation is used as the hidden representation of each SMILES. Whereby, SBERT introduces an additional intermediate fully connected layer to project word embeddings to a lower-dimensional space, which is followed by mean pooling.

6.1.3 Evaluation Metrics

QSAR

For the QSAR classification tasks, we use *Area Under The Curve of the Receiver Operating Characteristic* (AUROC) as specified in the MoleculeNet benchmark. The AUROC score essentially transforms the ROC curve into a single metric that depicts the model performance at various thresholds simultaneously. Importantly, a score of 1.0 on the AUROC metric entails perfect performance, whereby a 0.5 score entails model performance is equivalent to a random guess of the target label.

Virtual Screening

We use two metrics for the evaluation of language models-based molecular fingerprints. The first *Virtual Screening* evaluation metric corresponds to the AUROC score. Second, the *Boltzmann-Enhanced Discrimination of ROC* (BEDROC) [Truchon and Bayly, 2007] with (α = 20) which evaluates the early enrichment whilst allocating higher weights during retrieval to top $\alpha\%$ ranked SMILES with respect to Boltzmann distribution.

6.2 SMILES Representation Learning Results

To evaluate learned SMILES fingerprints, we evaluate the pre-trained language models on the *Virtual Screening* and QSAR downstream tasks. Furthermore, we directly address the research questions (RQs) in the experiments, each of which corresponds to the subsections from Sec. 6.2.1 onwards.

6.2.1 Encoder-based Contrastive Domain Adaptation (RQ1)

We evaluate the encoder-based SMILES language models as described in Sec. 5.1.1 against baseline elaborated in Sec. 6.1.1 on QSAR and *Virtual Screening* tasks. For both tasks, we compare the existing pre-training approaches to transfer learning with our proposed domain adaptation approaches.

QSAR

For downstream QSAR drug discovery tasks, we additionally also compare the effects of domain adaptation with another unlabeled proxy dataset. In our experiments, we use the MUV dataset from the *MoleculeNet* benchmark as the proxy dataset. We evaluate the models on four of the *MoleculeNet* benchmark datasets, as shown in Tab. 4.1.

	Adaptation	**BBBP**	**BACE**	**Tox21**	**ClinTox**
MLM BERT (Baseline)	–	0.686 ± 0.036	0.803 ± 0.034	0.711 ± 0.002	0.983 ± 0.018
MTR BERT	–	0.767 ± 0.032	0.796 ± 0.035	0.723 ± 0.002	0.965 ± 0.013
SBERT	P	0.713 ± 0.035	0.767 ± 0.037	0.690 ± 0.003	0.986 ± 0.0174
	S	0.717 ± 0.035	**0.820 ± 0.033**	0.707 ± 0.002	**0.992 ± 0.011**
CBERT	P	0.702 ± 0.036	0.790 ± 0.035	0.709 ± 0.003	0.979 ± 0.020
	S	0.729 ± 0.034	0.802 ± 0.034	0.705 ± 0.003	0.986 ± 0.016
MLM CBERT	P	0.702 ± 0.036	0.790 ± 0.035	0.709 ± 0.003	0.966 ± 0.026
	S	0.738 ± 0.034	0.807 ± 0.034	0.714 ± 0.002	0.990 ± 0.013
MTR CBERT	P	0.764 ± 0.033	0.774 ± 0.037	0.723 ± 0.002	0.972 ± 0.025
	S	0.720 ± 0.035	0.799 ± 0.035	0.716 ± 0.003	0.980 ± 0.019
MLM MTR BERT	P	0.765 ± 0.032	0.790 ± 0.035	0.723 ± 0.002	0.945 ± 0.036
	S	0.701 ± 0.036	0.739 ± 0.039	0.722 ± 0.002	0.970 ± 0.024
MLM MTR CBERT	P	**0.774 ± 0.032**	0.806 ± 0.034	**0.729 ± 0.002**	0.930 ± 0.041
	S	0.710 ± 0.035	0.800 ± 0.035	0.721 ± 0.003	0.960 ± 0.029

Table 6.2: Results for the MoleculeNet classification datasets on AUROC (↑) metric. "P" in the Adaption column indicates baseline model adaption on a proxy dataset (MUV), whereby, "S" indicates the same dataset as the target dataset.

On all datasets, CL-based domain adaptation when coupled with MLM and MTR objectives outperforms other objectives, as shown in Tab. 6.2. For the BBBP and Tox21 datasets, the best results are observed when all pre-training objectives of MLM, MTR, and CL are combined on the proxy dataset. Whereby, for BACE and ClinTox, SBERT-based domain adaptation on the target datasets appears to be most effective. Hence highlighting the quintessential importance of intermediate contrastive domain adaptation prior to fine-tuning.

Virtual Screening

For the *Virtual Screening* task, we use the RDKit benchmarking platform. The evaluation setting and the details about the datasets are described in detail in Sec. 4.1.2. Importantly, the objective of this experiment is to evaluate the robustness and fidelity of representations of our proposed domain-adapted language models.

	AUROC (↑)	BEDROC (↑)
ECFC4 (Baseline)	0.603 ± 0.056	0.170 ± 0.079
MLM BERT (Baseline)	0.615 ± 0.108	0.225 ± 0.102
MTR BERT	0.621 ± 0.121	0.262 ± 0.113
SBERT	0.673 ± 0.086	0.274 ± 0.108
CBERT	0.697 ± 0.091	0.270 ± 0.109
MLM CBERT	**0.708** ± **0.093**	0.281 ± 0.112
MTR CBERT	0.668 ± 0.111	0.285 ± 0.116
MLM MTR BERT	0.671± 0.107	**0.286** ± **0.114**
MLM MTR CBERT	0.666 ± 0.106	0.279 ± 0.113

Table 6.3: Results of Encoder-based approaches for *Virtual Screening* with the RDKit benchmarking platform.

Identical to the QSAR results on *Virtual Screening* the contrastive approaches demonstrate the most competitive results. Here, CBERT-based domain adaptation on MUV proves to be the most promising domain adaptation technique for the *Virtual Screening* task. Thereby, the results show that injecting enumeration awareness into the learned fingerprints using contrastive learning aids in the identification of true active molecules given a query molecule.

6.2.2 Denoising-based Encoder-decoder Canonicalization BART (RQ2)

Similar to the encoder-based language models, here we evaluate our proposed encoder-decoder denoising approach based on BART architecture as introduced in Sec. 5.1.2. The evaluation settings for both tasks are identical to encoder-based language models as explained in Sec. 6.2.1.

QSAR

For the evaluation of encoder-decoder-based models on QSAR downstream tasks, we contrast the influence of various stochastic noising schemes as described in Sec. 5.1.2. The overall evaluation setting remains identical to the one for encoder-based models as elaborated in Sec. 6.2.1.

	Noise	Adaptation	BBBP	BACE	Tox21	ClinTox
MLM BERT (Baseline)	–	–	0.686 ± 0.036	0.803 ± 0.034	**0.711 ± 0.002**	0.983 ± 0.018
MBART	Masking	–	0.704 ± 0.036	0.785 ± 0.036	0.687 ± 0.003	0.989 ± 0.013
PBART	Enumeration	–	0.721 ± 0.035	0.802 ± 0.034	0.707 ± 0.002	**0.990 ± 0.013**
MPBART	Masking + Enumeration	P	**0.745 ± 0.034**	0.812 ± 0.034	0.702 ± 0.003	0.989 ± 0.013
		S	0.731 ± 0.034	**0.813 ± 0.033**	0.701 ± 0.003	0.952 ± 0.031

Table 6.4: Results for the MoleculeNet classification datasets on AUROC (↑) metric. “P” in the Adaption column indicates baseline model adaption on the proxy dataset (MUV), whereby, “S” indicates the same dataset as the target dataset.

The results indicate that combining the noising schemes of masking and enumerations enable the models to learn more granular semantic knowledge of enumerations. Thus, the performance of BART models pre-trained with a combined noising scheme was better than single noising schemes. Although, denoising-based models performed reasonably on all QSAR datasets, however, they fail to outperform the best encoder-based domain adaption approaches on any of the datasets, as shown in Tab. 6.4.

Virtual Screening

In Tab. 6.5, we show the performance of denoising-based canonicalization language models on the *Virtual Screening*. Here, the evaluation setting is identical to the *Virtual Screening* evaluation task described Sec. 6.2.1 as well. Similarly, this experiment seeks to establish the expressiveness of representations learned by encoder-decoder-based canonicalization language models.

	Noise	AUROC (↑)	BEDROC (↑)
ECFC4 (Baseline)	–	0.603 ± 0.056	0.170 ± 0.079
MLM BERT (Baseline)	–	0.615 ± 0.108	0.225 ± 0.102
MTR BERT	–	0.621 ± 0.121	0.262 ± 0.113
MBART	Masking	0.615 ± 0.110	0.214 ± 0.099
PBART	Enumeration	0.646 ± 0.109	0.256 ± 0.110
MPBART	Masking + Enumeration	**0.660 ± 0.104**	**0.263 ± 0.111**

Table 6.5: Results of Encoder-decoder BART models for *Virtual Screening* with the RDKit benchmarking platform.

We observe that similar to the denoising-based evaluation of the QSAR tasks here as well, combined noising schemes outperform other noising schemes. Additionally, for *Virtual Screening* task as well, the contrastive encoder-based models’ performance stays superior as compared to the denoising-based encoder-decoder language models. Thereby indicating adopting encoder-based contrastive learning domain-adaptation is more beneficial for the downstream task of *Virtual Screening* in comparison to denoising-based encoder-decoder models.

6.2.3 Comparison to Others (RQ3)

We compare the performance of our proposed approaches to some of the most relevant existing molecular transformers-based language models. To be more specific, we juxtapose the results of our approaches to others on QSAR *MoleculeNet* datasets. The results shown in Tab. 6.6 demonstrate that our approaches consistently outperform the other approaches. Thereby, demonstrating that CL-based domain adaptation is beneficial for downstream QSAR tasks.

	BBBP	**BACE**	**Tox21**	**ClinTox**
SMILES Transformer [Honda et al., 2019]	0.684 ± 0.036	0.752 ± 0.039	0.695 ± 0.003	0.987 ± 0.016
ChemBERTa [Chithrananda et al., 2020]	0.643 ± 0.037	–	–	–
ChemBERTa-2 [Ahmad et al., 2022]	0.742 ± 0.034	0.799 ± 0.035	–	0.601 ± 0.040
Ours	**0.774 ± 0.032**	**0.820 ± 0.033**	**0.729 ± 0.002**	**0.992 ± 0.011**

Table 6.6: Results for the MoleculeNet classification datasets on AUROC (↑) metric compared to comparative approaches. We re-compute AUROC scores for **SMILES Transformer** [Honda et al., 2019] on scaffold splits as the original paper computes them on synthetic splits.

6.2.4 Evaluation of Enumeration-aware Representations (RQ4)

Here, we seek to evaluate which enumeration-aware pre-training method encodes the most SMILES enumeration knowledge into the language model representations. Precisely, we evaluate enumeration awareness in the following retrieval setting.

Let S be the canonical SMILE of a molecule. We enumerate S for m times to generate the set E. In our experiments, we use the value of $m = 1000$ in order to exhaust all valid enumerations of S. For the retrieval task, all the canonical SMILES in a dataset are used as queries Q. Whereby, the collection of Q for all canonical SMILES composes the document collection D. More formally as given below:

$$D = \bigcup_{i=1}^{N} E_i$$

Here, N is the size of Q and E_i is the set of all enumerations of i^{th} SMILES in Q. The evaluation process involves retrieving all enumerations of a canonical SMILES ($\forall e \in E$) from D for a given canonical SMILES as query $q \in Q$. We evaluate the retrieved SMILES on metrics including *Normalized Discounted Cumulative Gain* (NDCG)@K, Precision@K, Recall@K. Here, K corresponds to the number of retrieved SMILES for a given query. The evaluation results for mentioned metrics are presented in Fig. 6.1, Fig. 6.2, and Fig. 6.3 for the earlier mentioned metrics respectively.

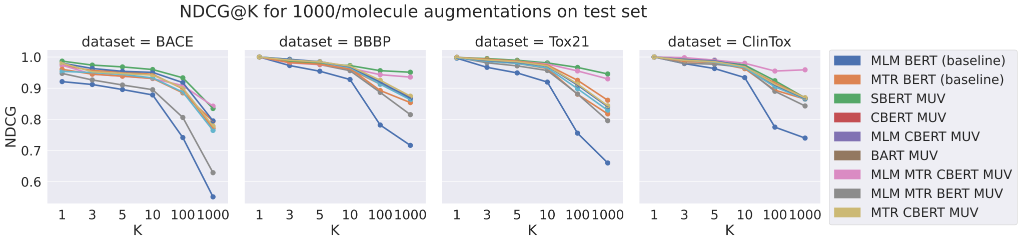

Figure 6.1: NDCG@K for SMILES retrieval on the MoleculeNet benchmark.

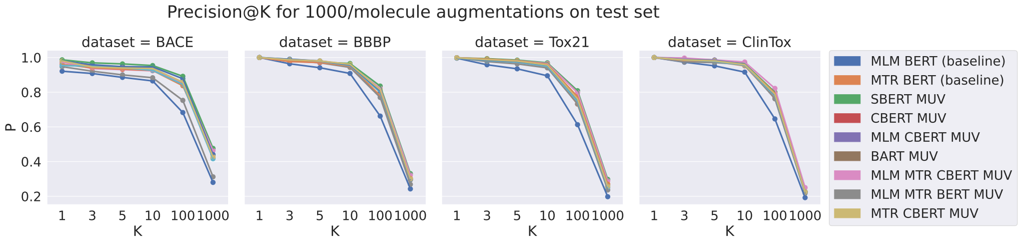

Figure 6.2: Precision@K for SMILES retrieval on the MoleculeNet benchmark.

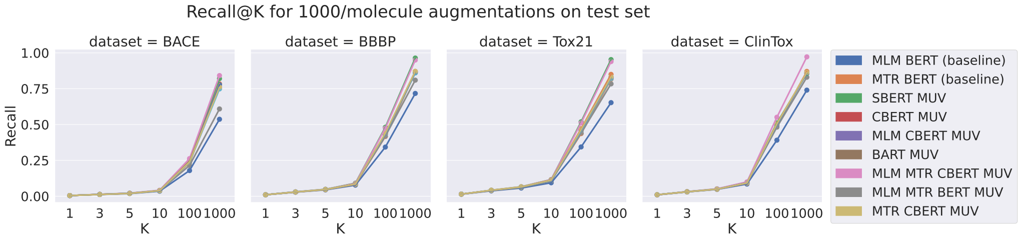

Figure 6.3: Recall@K for SMILES retrieval on the MoleculeNet benchmark.

From the retrieval experiments as shown above, we observe that CL-based domain-adapted language models outperform others. More specifically, CBERT and MLM CBERT consistently demonstrate superior enumeration retrieval capabilities on all metrics. For NDCG@K (Fig. 6.1), and Precision@K (Fig. 6.2) the retrieval performance of CBERT and MLM CBERT degrades the least for increasing values of K. Identically, for Recall@K (Fig. 6.3) the same approaches demonstrate the most performance gain as the values of K increase. In conclusion, the CL-based intermediate pre-training indeed embeds most enumeration-based knowledge into pre-trained molecular language models.

6.3 Enumeration-aware Semi-supervised Learning Results

For enumeration-aware SESL evaluation, we conduct experiments to investigate whether SESL approaches coupled with enumeration-aware pre-trained language models can alleviate label scarcity. Importantly, we extend enumeration-aware BERT-like language models by fine-tuning them with SESL approaches.

6.3.1 Semi-supervised Learning for Small Molecular Datasets (RQ5)

Our experiments, primarily focus on mitigating the low-data scenarios for QSAR-based downstream tasks. Furthermore, we use identical datasets from supervised learning approaches including BBBP, BACE, ClinTox, and Tox21. Notably, for ClinTox and Tox21 we use the most representative subtask **CT_TOX** and **SR-p53** respectively.

To evaluate our proposed approaches on a small molecular dataset, we synthetically simulate low-resource dataset scenarios. Specifically, by drawing a small random sample of size SPC for each class. Here, SPC corresponds to the number of samples per class with values including $\{50, 100, 150, 200, 250\}$. Furthermore, in order to get conclusive results, we repeat this process for $T = 20$ trials for each value of SPC. We present the results on BACE, BBBP, Tox21, and ClinTox datasets in Tab. 6.7, Tab. 6.8, Tab. 6.9, and Tab. 6.10 respectively.

	SPC@50	SPC@100	SPC@150	SPC@200	SPC@250
MLM BERT (Baseline)	0.669 ± 0.011	0.73 ± 0.012	0.745 ± 0.009	0.767 ± 0.009	0.786 ± 0.009
MTR BERT	0.671 ± 0.015	0.728 ± 0.010	0.728 ± 0.011	0.738 ± 0.007	0.751 ± 0.010
Pseudo-label	0.671 ± 0.013	0.764 ± 0.005	0.738 ± 0.010	0.745 ± 0.010	0.745 ± 0.006
Co-training	**0.763 ± 0.013**	**0.770 ± 0.016**	**0.765 ± 0.012**	**0.768 ± 0.024**	**0.791 ± 0.017**

Table 6.7: Results of Semi-supervised learning approaches on low-data **BACE dataset** using the AUROC (↑) metric. $SPC@N$ corresponds to the number of samples per class in the training dataset.

	$SPC@50$	$SPC@100$	$SPC@150$	$SPC@200$	$SPC@250$
MLM BERT (Baseline)	0.684 ± 0.0.007	**0.698 ± 0.007**	0.688 ± 0.008	**0.713 ± 0.008**	0.702 ± 0.009
MTR BERT	0.665 ± 0.009	0.697 ± 0.008	**0.702 ± 0.009**	0.703 ± 0.010	**0.728 ± 0.006**
Pseudo-label	0.661 ± 0.012	0.686 ± 0.008	0.692 ± 0.010	0.693 ± 0.008	0.694 ± 0.009
Co-training	**0.686 ± 0.013**	0.683 ± 0.015	0.700 ± 0.003	0.700 ± 0.006	0.705 ± 0.007

Table 6.8: Results of Semi-supervised learning approaches on low-data **BBBP dataset** using the AUROC (↑) metric. $SPC@N$ corresponds to the number of samples per class in the training dataset.

	$SPC@50$	$SPC@100$	$SPC@150$	$SPC@200$	$SPC@250$
MLM BERT (Baseline)	0.632 ± 0.010	0.642 ± 0.008	0.652 ± 0.007	0.670 ± 0.005	0.670 ± 0.004
MTR BERT	0.627 ± 0.007	0.639 ± 0.008	0.641 ± 0.007	0.650 ± 0.007	0.655 ± 0.005
Pseudo-label	**0.638 ± 0.009**	**0.768 ± 0.015**	**0.803 ± 0.012**	**0.793 ± 0.006**	**0.785 ± 0.006**
Co-training	0.627 ± 0.003	0.694 ± 0.004	0.694 ± 0.004	0.696 ± 0.003	0.701 ± 0.002

Table 6.9: Results of Semi-supervised learning approaches on low-data **Tox21 dataset** using the AUROC (↑) metric. $SPC@N$ corresponds to the number of samples per class in the training dataset.

	$SPC@50$	$SPC@100$	$SPC@150$	$SPC@200$	$SPC@250$
MLM BERT (Baseline)	0.955 ± 0.011	0.952 ± 0.012	0.951 ± 0.009	0.952 ± 0.009	0.958 ± 0.009
MTR BERT	0.954 ± 0.015	**0.956 ± 0.010**	**0.960 ± 0.011**	**0.957 ± 0.007**	**0.961 ± 0.010**
Pseudo-label	**0.957 ± 0.017**	0.953 ± 0.014	0.952 ± 0.015	0.952 ± 0.014	0.957 ± 0.013
Co-training	0.951 ± 0.008	0.953 ± 0.009	0.949 ± 0.009	0.948 ± 0.005	0.952 ± 0.004

Table 6.10: Results of Semi-supervised learning approaches on low-data **ClinTox dataset** using the AUROC (↑) metric. $SPC@N$ corresponds to the number of samples per class in the training dataset.

The results indicate that the SESL-based co-training and pseudo-label approaches significantly improve the performance of models trained on the BACE and Tox21 datasets. Furthermore, including additional unlabeled data can potentially lead to even better results. Hence, including SMILES data points with similar latent representation or physicochemical properties to increase the size of the unlabeled dataset can be a promising direction for future work.

6.4 Application: HIPS Dataset Results

We also evaluate our proposed approaches on a real-world low-data molecular dataset by HIPS for cytotoxicity, as described in Sec. 4.1.5. The results using SESL-based fine-tuning are shown below:

	AUROC (↑)
MLM BERT (Baseline)	0.838 ± 0.002
MTR BERT	0.831 ± 0.002
Pseudo-label	**0.869** ± **0.003**
Co-training	0.836 ± 0.002

Table 6.11: Results on the low-data HIPS dataset for cytotoxicity classification. For cell line HepG2 @ 100 µM.

Similar to the MoleculeNet benchmark, for the HIPS dataset we also observe that SESL fine-tuning such as the one based on pseudo-label considerably improves the model performance on the downstream low-data QSAR task.

6.5 Summary

In this chapter, we presented the results of the experiments to evaluate our proposed approaches. Furthermore, we systemically evaluated each research question associated with our proposed methods. The results indicate that incorporating enumeration awareness into molecular fingerprints can lead to significant improvement on downstream tasks such as *QSAR* and *Virtual Screening*. Furthermore, SSL-based pre-training approaches augmented with contrastive learning can best incorporate the enumeration knowledge into learned fingerprints. Finally, we also demonstrated that replacing fully supervised fine-tuning of language models with semi-supervised learning can help mitigate low-resource scenarios. Importantly, our evaluation process encompasses both the MoleculeNet benchmark and the real-world low-resource HIPS dataset.

Chapter 7
Conclusion

7.1 Summary

In this work, the primary goal is to learn high-fidelity and expressive molecular fingerprints from SMILES-based molecular datasets. Furthermore, our work serves as preliminary research on improving the generalization capabilities of pre-trained molecular language models for datasets with label scarcity problems. While the existing molecular language model pre-training regimes perform reasonably well. The SMILES-based MLM pre-training fails to take into account the enumeration knowledge, which is an essential property of molecular SMILES representation. Thus, the learned fingerprints belonging to the same molecular are prone to be distant in the molecular latent space. This results in performance degradation on downstream tasks such as *QSAR* and *Virtual Screening*.

To mitigate the absence of enumeration knowledge in the learned molecular fingerprints, we introduce two alternative pre-training objectives for transformer-based encoder-only and encoder-decoder architectures respectively. For BERT-like encoder-based architecture, we propose intermediate multi-task pre-training with contrastive learning called domain adaption. The contrastive learning pre-training not only infuses enumeration knowledge into the molecular fingerprints but also improves their alignment and distribution in the latent space. Whereby, for BART-like encoder-decoder architecture, we set the pre-training objective to learn a canonicalization function by denoising the corrupted input SMILES with various stochastic input noising functions.

Importantly, the acquisition of domain-specific datasets with specific protein affinities and properties is not feasible in most real-world molecular drug discovery settings. Whereby, the fine-tuning of the pre-trained language models requires the downstream dataset to be of reasonable size. Hence, the performance of the fine-tuned language model on small labeled datasets suffers. In our work, we propose to alleviate low-data settings for molecular language models by altering the fine-tuning process. Precisely, we propose semi-supervised learning approaches which can leverage additional unlabeled examples from the same distribution, yielding performance improvements on the downstream tasks such as *QSAR*.

In summary, our research project makes the following two key contributions:

- **Learning High-fidelity Molecular Fingerprints:** Modifying the pre-training regimes of molecular language models to infuse enumeration property into learned fingerprints. Thereby, enabling the fingerprints of the same molecule to be close in the molecular latent space.
- **Generalization on Low-data Scenarios:** Replacing fully supervised fine-tuning of language models with semi-supervised learning approaches. Hence, enabling generalization on datasets where further label acquisition is costly and non-trivial.

Our Enumeration-aware transformers improve over the baseline MLM-based language models on downstream drug discovery tasks. The evaluation process is based on two molecular downstream drug discovery tasks of *QSAR* and *Virtual Screening*. For QSAR, we evaluate both the encoder-only and encoder-decoder variants on four of the MoleculeNet datasets. These datasets include Tox21, ClinTox, BACE, and BBBP. Results indicate that our proposed multi-task domain adaptive pre-training achieves a performance gain of up to 9% on the AUROC metric. We also show that domain adaptation on a related proxy dataset instead of the downstream dataset also enhances the performance of the downstream model. For *Virtual Screening*, we use RDKit's benchmarking platform for evaluation purposes and observe up to 10% and 5% improvement on the AUROC and BEDROC20 metrics respectively. Lastly, we also empirically show that contrastive domain adaptation encodes the most enumeration knowledge into learned fingerprints.

For low-data scenarios, we evaluate our proposed semi-supervised methods on molecular datasets with label scarcity by synthetically simulating low-data settings on the MoleculeNet datasets as well as on a real-world low-resource dataset by the HIPS. For the synthetic low-data setting, we sample a fixed number of samples from each class and train the downstream model in a semi-supervised manner. We repeat the training and evaluation process for each dataset size for 20 trials to get conclusive results. Finally, our results show that semi-supervised learning improves performance on the AUROC metric by as much as 11%. Synonymously, we show that fine-tuning with the semi-supervised pseudo-label approach improves the AUROC score by $\sim$3% on the HIPS dataset. Thereby, demonstrating enumeration-aware semi-supervised fine-tuning as a promising direction for dealing with the scarcity of labels in future work.

7.2 Future Work

Our work on low-resource datasets only uses the molecules from the same distribution of the downstream dataset. Hence, this limits the application of the semi-supervised learning paradigm if a sufficient amount of unlabeled data from the same distribution is unavailable. This can be mitigated by including unlabeled molecules from other datasets which have similar underlying physicochemical properties or the fingerprints of the labeled molecules. Hence, evaluating such an acquisition method for unlabeled data can be an important facet of future work. Additionally, combining the labeled low-resource dataset with a reasonably large publicly available labeled dataset from the same domain can also be an alternative direction for future work.

List of Figures

List of Tables

Bibliography

[Ahmad et al., 2022] Ahmad, W., Simon, E., Chithrananda, S., Grand, G., and Ramsundar, B. (2022). Chemberta-2: Towards chemical foundation models.

[Amari, 1993] Amari, S.-i. (1993). Backpropagation and stochastic gradient descent method. *Neurocomputing*, 5(4-5):185–196.

[Arús-Pous et al., 2019] Arús-Pous, J., Johansson, S. V., Prykhodko, O., Bjerrum, E. J., Tyrchan, C., Reymond, J.-L., Chen, H., and Engkvist, O. (2019). Randomized smiles strings improve the quality of molecular generative models. *Journal of cheminformatics*, 11(1):1–13.

[Bahdanau et al., 2014] Bahdanau, D., Cho, K., and Bengio, Y. (2014). Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473*.

[Bjerrum and Sattarov, 2018a] Bjerrum, E. and Sattarov, B. (2018a). Applications of machine learning in drug discovery and development.

[Bjerrum, 2017] Bjerrum, E. J. (2017). Smiles enumeration as data augmentation for neural network modeling of molecules. *arXiv preprint arXiv:1703.07076*.

[Bjerrum and Sattarov, 2018b] Bjerrum, E. J. and Sattarov, B. (2018b). Improving chemical autoencoder latent space and molecular de novo generation diversity with heteroencoders. *Biomolecules*, 8(4):131.

[Blum and Mitchell, 1998] Blum, A. and Mitchell, T. (1998). Combining labeled and unlabeled data with co-training. In *Proceedings of the eleventh annual conference on Computational learning theory*, pages 92–100.

[Brown et al., 2019a] Brown, N., Fiscato, M., Segler, M. H., and Vaucher, A. C. (2019a). Guacamol: Benchmarking models for de novo molecular design. *Journal of Chemical Information and Modeling*, 59(3):1096–1108. PMID: 30887799.

[Brown et al., 2019b] Brown, N., Fiscato, M., Segler, M. H., and Vaucher, A. C. (2019b). Guacamol: benchmarking models for de novo molecular design. *Journal of chemical information and modeling*, 59(3):1096–1108.

[Cer et al., 2018] Cer, D., Yang, Y., Kong, S.-y., Hua, N., Limtiaco, N., John, R. S., Constant, N., Guajardo-Cespedes, M., Yuan, S., Tar, C., et al. (2018). Universal sentence encoder. *arXiv preprint arXiv:1803.11175*.

[Chen and Tseng, 2021] Chen, J.-H. and Tseng, Y. J. (2021). Different molecular enumeration influences in deep learning: an example using aqueous solubility. *Briefings in Bioinformatics*, 22(3):bbaa092.

[Chen et al., 2020] Chen, T., Kornblith, S., Norouzi, M., and Hinton, G. (2020). A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pages 1597–1607. PMLR.

[Chen et al., 2017] Chen, T., Sun, Y., Shi, Y., and Hong, L. (2017). On sampling strategies for neural network-based collaborative filtering. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 767–776.

[Chithrananda et al., 2020] Chithrananda, S., Grand, G., and Ramsundar, B. (2020). Chemberta: Large-scale self-supervised pretraining for molecular property prediction.

[Chopra et al., 2005] Chopra, S., Hadsell, R., and LeCun, Y. (2005). Learning a similarity metric discriminatively, with application to face verification. In *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)*, volume 1, pages 539–546. IEEE.

[Chung et al., 2014] Chung, J., Gulcehre, C., Cho, K., and Bengio, Y. (2014). Empirical evaluation of gated recurrent neural networks on sequence modeling. *arXiv preprint arXiv:1412.3555*.

[Cordero, 2021] Cordero, P. (2021). The four paths to molecular machine learning.

[Cortes and Vapnik, 1995] Cortes, C. and Vapnik, V. (1995). Support-vector networks. *Machine learning*, 20:273–297.

[Deshpande, 2022] Deshpande, A. (2022). Stereokg: A data-driven knowledge graph construction for cultural knowledge and stereotypes.

[Devlin et al., 2019] Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2019). 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)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

[Duvenaud et al., 2015] Duvenaud, D. K., Maclaurin, D., Iparraguirre, J., Bombarell, R., Hirzel, T., Aspuru-Guzik, A., and Adams, R. P. (2015). Convolutional networks on graphs for learning molecular fingerprints. *Advances in neural information processing systems*, 28.

[Ertl et al., 2017] Ertl, P., Lewis, R., Martin, E., and Polyakov, V. (2017). In silico generation of novel, drug-like chemical matter using the lstm neural network. *arXiv preprint arXiv:1712.07449*.

[Ethayarajh, 2019a] Ethayarajh, K. (2019a). How contextual are contextualized word representations? comparing the geometry of bert, elmo, and gpt-2 embeddings. *arXiv preprint arXiv:1909.00512*.

[Ethayarajh, 2019b] Ethayarajh, K. (2019b). How contextual are contextualized word representations? Comparing the geometry of BERT, ELMo, and GPT-2 embeddings. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 55–65, Hong Kong, China. Association for Computational Linguistics.

[Fabian et al., 2020] Fabian, B., Edlich, T., Gaspar, H., Segler, M., Meyers, J., Fiscato, M., and Ahmed, M. (2020). Molecular representation learning with language models and domain-relevant auxiliary tasks.

[Gao et al., 2021] Gao, T., Yao, X., and Chen, D. (2021). SimCSE: Simple contrastive learning of sentence embeddings. In *Empirical Methods in Natural Language Processing (EMNLP)*.

[Gaulton et al., 2017] Gaulton, A., Hersey, A., Nowotka, M., Bento, A. P., Chambers, J., Mendez, D., Mutowo, P., Atkinson, F., Bellis, L. J., Cibrián-Uhalte, E., et al. (2017). The chembl database in 2017. *Nucleic acids research*, 45(D1):D945–D954.

[Gómez-Bombarelli et al., 2018] Gómez-Bombarelli, R., Wei, J. N., Duvenaud, D., Hernández-Lobato, J. M., Sánchez-Lengeling, B., Sheberla, D., Aguilera-Iparraguirre, J., Hirzel, T. D., Adams, R. P., and Aspuru-Guzik, A. (2018). Automatic chemical design using a data-driven continuous representation of molecules. *ACS central science*, 4(2):268–276.

[Guo et al., 2022] Guo, Z., Sharma, P., Martinez, A., Du, L., and Abraham, R. (2022). Multilingual molecular representation learning via contrastive pre-training. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3441–3453.

[Henderson et al., 2017] Henderson, M., Al-Rfou, R., Strope, B., Sung, Y.-H., Lukács, L., Guo, R., Kumar, S., Miklos, B., and Kurzweil, R. (2017). Efficient natural language response suggestion for smart reply. *arXiv preprint arXiv:1705.00652*.

[Hendrycks and Gimpel, 2016] Hendrycks, D. and Gimpel, K. (2016). Gaussian error linear units (gelus). *arXiv preprint arXiv:1606.08415*.

[Hochreiter and Schmidhuber, 1997] Hochreiter, S. and Schmidhuber, J. (1997). Long short-term memory. *Neural computation*, 9(8):1735–1780.

[Honda et al., 2019] Honda, S., Shi, S., and Ueda, H. R. (2019). Smiles transformer: Pre-trained molecular fingerprint for low data drug discovery.

[Irwin et al., 2020] Irwin, J. J., Tang, K. G., Young, J., Dandarchuluun, C., Wong, B. R., Khurelbaatar, M., Moroz, Y. S., Mayfield, J., and Sayle, R. A. (2020). Zinc20—a free ultralarge-scale chemical database for ligand discovery. *Journal of Chemical Information and Modeling*, 60(12):6065–6073. PMID: 33118813.

[Irwin et al., 2022] Irwin, R., Dimitriadis, S., He, J., and Bjerrum, E. J. (2022). Chemformer: a pre-trained transformer for computational chemistry. *Machine Learning: Science and Technology*, 3(1):015022.

[James, 2004] James, C. A. (2004). Daylight theory manual. *http://www. daylight. com/dayhtml/doc/theory/theory. toc. html*.

[Jastrzebski et al., 2016] Jastrzebski, S., Leśniak, D., and Czarnecki, W. M. (2016). Learning to smile (s). *arXiv preprint arXiv:1602.06289*.

[Kearnes et al., 2016] Kearnes, S., McCloskey, K., Berndl, M., Pande, V., and Riley, P. (2016). Molecular graph convolutions: moving beyond fingerprints. *Journal of computer-aided molecular design*, 30(8):595–608.

[Kingma and Welling, 2013] Kingma, D. P. and Welling, M. (2013). Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*.

[Kiros et al., 2015] Kiros, R., Zhu, Y., Salakhutdinov, R. R., Zemel, R., Urtasun, R., Torralba, A., and Fidler, S. (2015). Skip-thought vectors. *Advances in neural information processing systems*, 28.

[Klambauer et al., 2019] Klambauer, G., Hochreiter, S., and Rarey, M. (2019). Machine learning in drug discovery.

[Kolesnikov et al., 2021] Kolesnikov, A., Dosovitskiy, A., Weissenborn, D., Heigold, G., Uszkoreit, J., Beyer, L., Minderer, M., Dehghani, M., Houlsby, N., Gelly, S., Unterthiner, T., and Zhai, X. (2021). An image is worth 16x16 words: Transformers for image recognition at scale.

[Kusner et al., 2017] Kusner, M. J., Paige, B., and Hernández-Lobato, J. M. (2017). Grammar variational autoencoder. In *International conference on machine learning*, pages 1945–1954. PMLR.

[Lee et al., 2013] Lee, D.-H. et al. (2013). Pseudo-label: The simple and efficient semi-supervised learning method for deep neural networks. In *Workshop on challenges in representation learning, ICML*, volume 3, page 896.

[Lewis et al., 2019] Lewis, M., Liu, Y., Goyal, N., Ghazvininejad, M., Mohamed, A., Levy, O., Stoyanov, V., and Zettlemoyer, L. (2019). Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. *arXiv preprint arXiv:1910.13461*.

[Li et al., 2020a] Li, B., Zhou, H., He, J., Wang, M., Yang, Y., and Li, L. (2020a). On the sentence embeddings from pre-trained language models. *arXiv preprint arXiv:2011.05864*.

[Li et al., 2020b] Li, B., Zhou, H., He, J., Wang, M., Yang, Y., and Li, L. (2020b). On the sentence embeddings from pre-trained language models. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 9119–9130, Online. Association for Computational Linguistics.

[Li et al., 2022] Li, C., Wang, C., Sun, M., Zeng, Y., Yuan, Y., Gou, Q., Wang, G., Guo, Y., and Pu, X. (2022). Correlated rnn framework to quickly generate molecules with desired properties for energetic materials in the low data regime. *Journal of Chemical Information and Modeling*.

[Li and Jiang, 2021] Li, J. and Jiang, X. (2021). Mol-bert: An effective molecular representation with bert for molecular property prediction. *Wireless Communications and Mobile Computing*, 2021.

[Liu et al., 2017] Liu, B., Ramsundar, B., Kawthekar, P., Shi, J., Gomes, J., Luu Nguyen, Q., Ho, S., Sloane, J., Wender, P., and Pande, V. (2017). Retrosynthetic reaction prediction using neural sequence-to-sequence models. *ACS central science*, 3(10):1103–1113.

[Maziarka et al., 2020] Maziarka, L., Danel, T., Mucha, S., Rataj, K., Tabor, J., and Jastrzkebski, S. (2020). Molecule attention transformer. *arXiv preprint arXiv:2002.08264*.

[Mohs and Greig, 2017] Mohs, R. C. and Greig, N. H. (2017). Drug discovery and development: Role of basic biological research. *Alzheimer's & Dementia: Translational Research & Clinical Interventions*, 3(4):651–657.

[Morgan, 1965] Morgan, H. L. (1965). The generation of a unique machine description for chemical structures-a technique developed at chemical abstracts service. *Journal of chemical documentation*, 5(2):107–113.

[O'Boyle and Dalke, 2018] O'Boyle, N. and Dalke, A. (2018). Deepsmiles: an adaptation of smiles for use in machine-learning of chemical structures.

[Oord et al., 2018] Oord, A. v. d., Li, Y., and Vinyals, O. (2018). Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*.

[Payne et al., 2020] Payne, J., Srouji, M., Yap, D. A., and Kosaraju, V. (2020). Bert learns (and teaches) chemistry.

[Pinheiro et al., 2022] Pinheiro, G. A., Da Silva, J. L., and Quiles, M. G. (2022). Smiclr: Contrastive learning on multiple molecular representations for semisupervised and unsupervised representation learning. *Journal of Chemical Information and Modeling*.

[Reimers and Gurevych, 2019] Reimers, N. and Gurevych, I. (2019). Sentence-bert: Sentence embeddings using siamese bert-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.

[Reusch, 1999] Reusch, W. (1999). Virtual textbook of organic chemistry.

[Riniker and Landrum, 2013] Riniker, S. and Landrum, G. A. (2013). Open-source platform to benchmark fingerprints for ligand-based virtual screening. *Journal of cheminformatics*, 5(1):1–17.

[Rogers and Hahn, 2010] Rogers, D. and Hahn, M. (2010). Extended-connectivity fingerprints. *Journal of chemical information and modeling*, 50(5):742–754.

[Rohrer and Baumann, 2009] Rohrer, S. G. and Baumann, K. (2009). Maximum unbiased validation (muv) data sets for virtual screening based on pubchem bioactivity data. *Journal of chemical information and modeling*, 49(2):169–184.

[Schwaller et al., 2019] Schwaller, P., Laino, T., Gaudin, T., Bolgar, P., Hunter, C. A., Bekas, C., and Lee, A. A. (2019). Molecular transformer: a model for uncertainty-calibrated chemical reaction prediction. *ACS central science*, 5(9):1572–1583.

[Shorten and Khoshgoftaar, 2019] Shorten, C. and Khoshgoftaar, T. M. (2019). A survey on image data augmentation for deep learning. *Journal of big data*, 6(1):1–48.

[Shorten et al., 2021] Shorten, C., Khoshgoftaar, T. M., and Furht, B. (2021). Text data augmentation for deep learning. *Journal of big Data*, 8(1):1–34.

[Shrivastava and Kell, 2021] Shrivastava, A. D. and Kell, D. B. (2021). Fragnet, a contrastive learning-based transformer model for clustering, interpreting, visualizing, and navigating chemical space. *Molecules*, 26(7):2065.

[Skinnider et al., 2021] Skinnider, M. A., Stacey, R. G., Wishart, D. S., and Foster, L. J. (2021). Chemical language models enable navigation in sparsely populated chemical space. *Nature Machine Intelligence*, 3(9):759–770.

[Smith et al., 2018] Smith, J. S., Roitberg, A. E., and Isayev, O. (2018). Transforming computational drug discovery with machine learning and ai.

[Soliman, 2022] Soliman, H. (2022). Cross-domain neural entity linking.

[Subramanian et al., 2016] Subramanian, G., Ramsundar, B., Pande, V., and Denny, R. A. (2016). Computational modeling of β-secretase 1 (bace-1) inhibitors using ligand based approaches. *Journal of chemical information and modeling*, 56(10):1936–1949.

[Taylor and Nitschke, 2018] Taylor, L. and Nitschke, G. (2018). Improving deep learning with generic data augmentation. In *2018 IEEE Symposium Series on Computational Intelligence (SSCI)*, pages 1542–1547. IEEE.

[Tetko et al., 2019] Tetko, I. V., Karpov, P., Bruno, E., Kimber, T. B., and Godin, G. (2019). Augmentation is what you need! In *International Conference on Artificial Neural Networks*, pages 831–835. Springer.

[Truchon and Bayly, 2007] Truchon, J.-F. and Bayly, C. I. (2007). Evaluating virtual screening methods: good and bad metrics for the "early recognition" problem. *Journal of chemical information and modeling*, 47(2):488–508.

[Vamathevan et al., 2019] Vamathevan, J., Clark, D., Czodrowski, P., Dunham, I., Ferran, E., Lee, G., Li, B., Madabhushi, A., Shah, P., Spitzer, M., et al. (2019). Applications of machine learning in drug discovery and development. *Nature reviews Drug discovery*, 18(6):463–477.

[van Deursen et al., 2020] van Deursen, R., Ertl, P., Tetko, I. V., and Godin, G. (2020). Gen: highly efficient smiles explorer using autodidactic generative examination networks. *Journal of Cheminformatics*, 12(1):1–14.

[Vaswani et al., 2017] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., and Polosukhin, I. (2017). Attention is all you need. *Advances in neural information processing systems*, 30.

[Wang et al., 2019] Wang, S., Guo, Y., Wang, Y., Sun, H., and Huang, J. (2019). Smiles-bert: large scale unsupervised pre-training for molecular property prediction. In *Proceedings of the 10th ACM international conference on bioinformatics, computational biology and health informatics*, pages 429–436.

[Wang et al., 2021] Wang, Y., Wang, J., Cao, Z., and Farimani, A. B. (2021). Molclr: Molecular contrastive learning of representations via graph neural networks. *arXiv preprint arXiv:2102.10056*.

[Webel et al., 2020] Webel, H. E., Kimber, T. B., Radetzki, S., Neuenschwander, M., Nazaré, M., and Volkamer, A. (2020). Revealing cytotoxic substructures in molecules using deep learning. *Journal of computer-aided molecular design*, 34(7):731–746.

[Weininger, 1988] Weininger, D. (1988). Smiles, a chemical language and information system. 1. introduction to methodology and encoding rules. *Journal of Chemical Information and Computer Sciences*, 28(1):31–36.

[Weininger et al., 1989] Weininger, D., Weininger, A., and Weininger, J. L. (1989). Smiles. 2. algorithm for generation of unique smiles notation. *Journal of chemical information and computer sciences*, 29(2):97–101.

[Winter et al., 2019] Winter, R., Montanari, F., Noé, F., and Clevert, D.-A. (2019). Learning continuous and data-driven molecular descriptors by translating equivalent chemical representations. *Chemical science*, 10(6):1692–1701.

[Wu et al., 2022] Wu, Z., Jiang, D., Wang, J., Zhang, X., Du, H., Pan, L., Hsieh, C.-Y., Cao, D., and Hou, T. (2022). Knowledge-based bert: a method to extract molecular features like computational chemists. *Briefings in Bioinformatics*, 23(3):bbac131.

[Wu et al., 2018] Wu, Z., Ramsundar, B., Feinberg, E. N., Gomes, J., Geniesse, C., Pappu, A. S., Leswing, K., and Pande, V. (2018). Moleculenet: a benchmark for molecular machine learning. *Chemical science*, 9(2):513–530.

[Xu et al., 2017] Xu, Z., Wang, S., Zhu, F., and Huang, J. (2017). Seq2seq fingerprint: An unsupervised deep molecular embedding for drug discovery. In *Proceedings of the 8th ACM international conference on bioinformatics, computational biology, and health informatics*, pages 285–294.

[Yang et al., 2021a] Yang, X., Song, Z., King, I., and Xu, Z. (2021a). A survey on deep semi-supervised learning. *arXiv preprint arXiv:2103.00550*.

[Yang et al., 2018] Yang, Y., Yuan, S., Cer, D., Kong, S.-y., Constant, N., Pilar, P., Ge, H., Sung, Y.-H., Strope, B., and Kurzweil, R. (2018). Learning semantic textual similarity from conversations. *arXiv preprint arXiv:1804.07754*.

[Yang et al., 2021b] Yang, Z., Song, J., Yang, M., Yao, L., Zhang, J., Shi, H., Ji, X., Deng, Y., and Wang, X. (2021b). Cross-modal retrieval between 13c nmr spectra and structures for compound identification using deep contrastive learning. *Analytical Chemistry*, 93(50):16947–16955.

[Zhang et al., 2021a] Zhang, X.-C., Wu, C.-K., Yang, Z.-J., Wu, Z.-X., Yi, J.-C., Hsieh, C.-Y., Hou, T.-J., and Cao, D.-S. (2021a). Mg-bert: leveraging unsupervised atomic representation learning for molecular property prediction. *Briefings in bioinformatics*, 22(6):bbab152.

[Zhang et al., 2021b] Zhang, Y., Wang, L., Wang, X., Zhang, C., Ge, J., Tang, J., Su, A., and Duan, H. (2021b). Data augmentation and transfer learning strategies for reaction prediction in low chemical data regimes. *Organic Chemistry Frontiers*, 8(7):1415–1423.

Appendix A
Code Modules

The experimental and evaluation pipelines corresponding to this thesis are implemented in the form of multiple modules. This appendix chapter briefly highlights the functionality of each of those modules. The precise details about the usage of each module can be found on their respective web pages.

A.1 Enumeration-aware Molecular Transformers

This module contains the code for pre-training enumeration-aware transformers. It includes the implementation of training with contrastive learning alongside multi-task regression, and masked language modeling as pre-training objectives to inject enumeration knowledge into pre-trained language models. The corresponding code can be found at https://github.com/MoleculeTransformers/enumeration-aware-molecule-transformers

A.2 Fine-tuning on low-data via Semi-supervised Learning

This work package contains the code implementation that replaces fully supervised fine-tuning of molecular language models with semi-supervised learning methods including pseudo-label, and deep co-training to generalize language models in low-data scenarios. The corresponding code can be found at https://github.com/MoleculeTransformers/moleculenet-bert-ssl

A.3 RDKit Virtual Screening Benchmarking Platform for Transformers

We implemented the port of the RDKit Virtual Screening Benchmarking Platform[3] to evaluate enumeration-aware fingerprints on the *Virtual Screening* task. The corresponding code can be found at https://github.com/MoleculeTransformers/rdkit-benchmarking-platform-transformers

A.4 SMILES Featurizers

This code module allows obtaining enumeration-aware fingerprints from various out-of-the-box pre-trained molecule transformers. The corresponding pre-trained models and code implementation can be found at https://github.com/MoleculeTransformers/smiles-featurizers

A.5 SMILES Augment

This code module provides different stochastic and interpolation-oriented SMILES augmentation mechanisms to augment SMILES-based training datasets. The corresponding code can be found at https://github.com/MoleculeTransformers/smiles-augment

[3] `https://github.com/rdkit/benchmarking_platform`.