Advancing Machine Learning Methods for Small and Large Biological Datasets: Applications in Redox Potential Prediction, Enzyme Interaction Prediction, and Spectra Annotation

A dissertation submitted by

Apurva Kalia

in partial fulfillment of the requirements for the degree of
Doctor of Philosophy
in
Computer Science
Tufts University
February 2025

Adviser: Soha Hassoun

Abstract

Advancements in computational methods, especially with deep learning and AI, have driven breakthroughs in areas such as computer vision and natural language processing. This progress has sparked efforts to apply these powerful techniques to complex biological problems, such as drug design, protein engineering, and metabolic engineering. However, biological data presents unique challenges since it is frequently multi-modal, limited in size due to costly experimentation, and often noisy. These factors add complexity to feature selection and learning. Consequently, it is essential to develop computational methods tailored to the characteristics and quality of data available for each specific biological problem. This thesis addresses three biological problems, where one problem has a data-scarce scenario requiring feature selection and evaluation, while the other two problems have data-abundant scenarios enabling the use of deep learning for representation learning.

First, we address the challenge of predicting redox (reduction/oxidation) potential, which is a molecular property that is crucial for understanding electron transfer processes in chemical and biological systems. As there are currently no established ”best” method for creating redox models for small datasets, we explore the use of existing molecular features, including structural properties, molecular energies, and drug-like attributes, to develop Gaussian Process Regression (GPR) models for predicting redox potentials across three datasets of organic molecules. We demonstrate that combining these features yields higher predictive accuracy than when using each feature independently, and validate the model's performance on an experimental quinone dataset. Our results highlight the importance of integrating structural, electronic, and pharmacokinetic properties to capture critical redox characteristics, offering a recommended method for accurate redox potential prediction when dealing with smaller datasets. Next, we explore the problem of molecule-enzyme interaction prediction, a critical factor in understanding enzyme promiscuity and its implications for biocatalysis and evolutionary biology. We develop Contrastive Stratification for Interaction Prediction, CSI,

a novel model which partitions data by interaction features and applies contrastive learning to generate embeddings to maximize mutual information across congruent views. By incorporating multiple views on enzymatic reactions, CSI learns high-quality representations of enzymes and molecules, resulting in improved interaction predictions over methods that do not utilize such views. Finally, we address the problem of annotating spectra measured via untargeted mass spectrometry. Assigning chemical identities to the measured spectra is essential for identifying biomarkers and diagnosing diseases. Despite recent computational efforts, annotation rates remain low. We present Joint Embedding Space Technique for Ranking, JESTR, a deep-learning approach to embed both molecular structures and their spectra into a joint embedding space. Through contrastive learning, JESTR aligns molecular and spectral representations, ranking candidate molecules by cosine similarity to the query spectrum. JESTR improves annotation accuracy over state-of-the-art existing methods by avoiding the explicit reconstruction of molecules from spectra, or the construction of spectra from candidate molecules. Collectively, these projects illustrate the importance of exploiting data to derive high-quality features, achieved through either manual engineering or deep learning, in addressing critical biological challenges.

Acknowledgments

I would like to express my deepest gratitude to my wonderful and patient advisor - Professor Soha Hassoun. Soha is the reason that I had the gumption of attempting a PhD program in a field vastly different from the one in which I had spent 3 decades of my life. She guided me with immense patience while giving ample space to explore new ideas. Thank you Soha - I could not have done this without your help and support.

I would also like to thank my committee members for their guidance. Thanks are especially due to Dr Dilip Krishnan - his work on contrastive learning inspired the work on two of my projects. Professor Liping Liu introduced me to the wonderful world of Deep Neural Networks. I thank Professor Sameer Sonkusale and Professor Remco Chang to be on my committee and provide valuable insights and feedback on my research work and thesis.

My research colleagues in Hassoun Lab made this journey very enjoyable. I benefited a lot from the numerous discussions and brainstorming sessions with them. Thanks are also due to the Computer Science department at Tufts for affording me the flexibility that I needed because of the long break in my academic career.

Finally, and most importantly, I humbly, lovingly and gratefully acknowledge the phenomenal support of the three strong women in my life - my two wonderful daughters and my amazing wife Kirti. They encouraged my decision to give up a thriving corporate career and make this huge leap. My daughters taught me the basics of biology and chemistry and were the sounding boards for my strangest ideas.

This research is supported by NSF Award 1909536, NIGMS of the National Institutes of Health, Award R01GM132391 and Award R35GM148219, and Army Research Office, MURI program, contract W911NF2210239. Any views and conclusions contained herein are those of the author, and do not necessarily represent the official positions, express or implied, of the funders.

Table of Contents

List of Tables

List of Figures

Chapter 1 - Introduction

Computational methods have advanced rapidly, with deep learning and AI driving major breakthroughs in area such as computer vision and natural language processing. This success has fueled an interest in adapting these techniques to tackle problems in other fields, including biology. Areas such as drug development, enzyme engineering, and metabolic engineering, are poised to benefit from such computational advances. However, biological data introduces unique challenges. Biological data is often multi-modal, sparse, and noisy, therefore complicating feature selection and representation learning. This thesis investigates three significant Bioinformatics challenges. The first is molecular property prediction, focusing on predicting the redox potential of organic molecules. This problem is data scarce with insufficient data for representation learning, and hence requires a feature selection approach. The second challenge addresses the need for characterizing enzyme promiscuity by predicting molecule-enzyme interaction likelihood. The final challenge is the annotation of untargeted spectra, a critical step for identifying compounds in biological samples. The last two problems have data abundant contexts, and we leverage the data for learning high-quality representations. Overall, this thesis advocates for methods that are adaptable, capable of mitigating data sparsity, or harnessing the advantages of data abundance. The results obtained across the three challenges highlight that effective feature learning must be customized to meet the unique demands of each Bioinformatics problem.

1.1 High-throughput prediction of redox potential

Predicting molecular properties is crucial in understanding how a molecule interacts within a biological system. Molecular properties such as solubility, binding affinity, and stability can influence a molecule's efficacy, bioavailability, and safety [1]. Among these, reduction/oxidation (redox) potential is especially important, as it defines a molecule's ability to gain or lose electrons in metabolic reactions. Accurate prediction of redox potentials aids in the design of molecules tailored for specific enzyme interactions, optimizing catalytic efficiency in processes ranging from cellular metabolism to therapeutic development [2, 3]. Redox potential, the tendency of a chemical species to either acquire electrons (reduction) from or lose electrons (oxidation) to another electron donor or acceptor, is a key property that plays a crucial role in governing the direction and feasibility of electron-transfer based biochemical reactions [4, 5, 6]. For example, the electron transfer chain in cellular respiration depends on a series of redox reactions involving NADH, FADH2, ubiquinone, membrane protein complexes and a terminal electron acceptor. The redox potential differences between the components allow electrons to flow through the system, releasing energy to facilitate ATP synthesis [7]. In another example, the redox potential of the peptide glutathione allows reversible interactions with a variety of redox couples in the cell. Such interactions are critical in maintaining redox homeostasis within the cell [8]. Redox potential not only influences the thermodynamics of these critical biological processes, but also plays a crucial role in bioelectronic applications such as bioelectrocatalysis and bioelectronic sensing. In these applications, electroactive bacteria or redox active enzymes catalyze electron transfer to and from electrodes for purposes including analyte sensing, energy generation, and chemical synthesis [9, 10, 11, 12, 13, 14]. Biological energy transduction is most commonly mediated by quinones and is widely studied for its application in the pharmaceutical industry to combat pathogens and inflammation [15, 16]. Other applications where redox reactions play a key role include organic synthesis, flow batteries and solar cells [17, 18, 19].

Redox potential is primarily determined by the intrinsic electronic properties of the molecule, the chemical environment (pH, solvent), and external conditions (temperature, concentration), all of which influence the tendency of the molecule to undergo electron transfer. Experimentally, redox potential is measured using cyclic voltammetry [20], and the readout is measured under specific reaction conditions, such as temperature, pressure, and solvent. Experimental efforts can be costly and time-consuming, so computational methods for the accurate prediction of redox potential can provide advantages such as early-stage prediction to guide experimental workflows, high-throughput screening of compound libraries, and molecular-level insights within complex systems. As there are currently no standard methods for modeling redox potential for small datasets, we explore in this thesis the use of various molecular descriptors across three datasets to recommend a set of features appropriate for predicting redox potential for organic molecules based on a small set of experimental conditions.

1.2 Compound-protein interaction prediction

Exploring the diverse catalytic roles of enzymes is essential for progress in fields such as biochemistry, molecular biology, and synthetic biology. Though traditionally viewed as highly specific catalysts for individual substrates, it is now widely recognized that enzymes can act on a variety of substrates, including those outside their natural evolutionary roles. The inherent promiscuity of enzymes is valuable for protein engineering, as it enables the design of new enzymes for novel synthetic pathways, aids in the discovery of metabolic pathways in natural and engineered organisms, and supports the development of innovative therapeutic and industrial compounds. Despite significant strides in protein function annotation and protein-ligand interaction modeling, particularly with drug-ligand systems, the full catalytic potential of enzymes remains uncharacterized. Computational tools for predicting the likelihood of interactions between enzymes and molecules can fill these gaps, guiding applications across biology and biomedicine while reducing the need for resource-intensive

experimental validation. This thesis explores the stratification of enzymatic data to enhance the prediction of interaction likelihood.

1.3 Annotation of untargeted MS-MS spectra

Metabolomics is an impactful field of "omics" involving the measurement and analysis of *masses* of small molecules in biological samples [21]. Unlike larger molecular structures such as proteins, DNA, and RNA, small molecules have relatively low molecular weight, typically less than 1000 Da (Daltons). Small molecules are often products of metabolism, and hence referred to as metabolites. Importantly, small molecules play crucial roles in biological processes, serving as building blocks, intermediates, and regulators, and they can contribute to biomarker discovery, drug development, plant biology, nutrition, environmental health, and many applications.

The ability to use *untargeted metabolomics*, where masses of thousands of metabolites within a biological sample are detected, presents unprecedented opportunities to characterize the metabolome. Annotation, the process of assigning chemical structures to metabolomics measurements, however is riddled with uncertainty. Naïvely, one can presume the measured mass could be used to determine a metabolite's molecular structure. However, a particular molecular mass can map to possibly thousands of candidate molecular structures sharing the same chemical formula (e.g., there are 44,374 known molecular structures associated with $C_{12}H_{18}N_2O_2$).

With advances in mass spectrometry instrumentation, it is now possible to not only measure the mass of ionized molecules, but to also measure masses of ionized molecular fragments. Techniques such as combining liquid chromatography (LC) with mass spectrometry (MS) or combining two such analysis steps (tandem MS/MS) have now become dominant in metabolomics. The measured mass spectrum is a collection of peaks. Each peak is represented by its mass-to-charge (m/z) ratio, where the charge is known and is often +1 or -1, and a relative intensity. Even for an experienced analytical chemist, assigning a

chemical structure to LC-MS or MS/MS spectra is an unsolved problem as the spectrum provides a partial view on the measured molecule. Hence, annotating untargeted spectra is a challenging task. Importantly, computational methods can help address this challenge, where data within spectral libraries can be used for model training. In this thesis, we formulate the annotation problem anew, and propose a novel approach for spectra annotation problem that does not explicitly reconstruct the data label.

1.4 Thesis contribution

To address the problem of redox potential prediction for organic molecules, we present a systematic study to discern a generalizable method for predicting redox potential and validate our findings against a new experimental dataset. We select three redox datasets with measured experimental data for our analysis. Quinones are an important class of redox active organic molecules. Relevant experimental datasets are available for one electron reduction potentials for various types of quinones [22]. We use the 1,4-benzoquinone and 1,4-naphthoquinone datasets listed in Prince et al. [22]. The ROP313 dataset [23] is a widely used dataset consisting of 313 organic and organometallic molecules whose one electron reduction potentials have been measured in an aqueous solvent. A subset of these include 183 organic molecules labeled as the OROP dataset. Importantly, we test our model using an experimental dataset of redox potentials experimentally measured for a small quinone library [10]. While two-electron, two-proton reduction data is also available, we focus in this thesis on one-electron reduction because of the availability of diverse datasets for training, as well as experimental data for testing. Our proposed model explores the use of Morgan fingerprints, molecular descriptors as well as DFT energies to predict redox potentials. However, to further improve the performance of predictions, we propose a novel application of pharmacokinetic properties such as ADME (absorption, distribution, metabolism, excretion) [24] to this problem, as ADME properties were shown previously to correlate with redox activity [25, 26].

To predict the interaction likelihood between a compound and an enzyme, we model the problem as predicting the likelihood of two interacting objects over multiple views of the data. Predicting the likelihood of interaction between two objects (e.g., user-item, spectator-movie, author-paper, label-image, compound-protein, and other pairs) is a fundamental problem in Computer Science. Recommender systems, for example, utilize methods based on matrix-factorization to predict unknown interactions between users and items [27, 28]. In network graphs, link prediction methods can anticipate potential connections between two collaborators, or authors and papers [29]. Image captioning is achieved by recognizing objects within an image and characterizing interactions among them[30]. All these methods can also be applied to the problem of compound-protein interaction prediction. Across various tasks, the success of interaction prediction hinges on learned representations of the interacting objects, as high-quality representations capture key features of interest.

Multiple strategies have been developed to generate compressed representations of data [31, 32, 33]. Importantly, the availability of multi-modal data representing different aspects of the same object creates opportunities for multi-view learning techniques [34], which have proven to be a powerful way to learn representations, especially in the computer vision literature [35, 36]. Some such techniques attempt to minimize the distances between congruent (same-object) views, while others contrast congruent and non-congruent views of the data to push away embeddings of differing data points.

When addressing the interaction prediction problem, multi-modal representation learning can be applied on each object involved in the interaction. In this case, each object is embedded within its own latent space. In some tasks, deriving congruent data views is a common place task, e.g., image cropping, chrominance, and luminance for image-related tasks. However, in other cases, identifying congruent multi-views of data is challenging or non-trivial (e.g., drugs, disease, etc). To address this issue, and to further improve on representation learning for interactions, we use stratification (data partitioning) to generate multiple views of the data and to establish congruent and non-congruent views. Contrastive

learning methods can then be applied on the stratified data to enhance learning. We develop in this thesis a novel method for stratifying enzymatic data to enhance interaction prediction.

Finally, to address the problem of untargeted spectra annotation, we focus on the problem of ranking candidates molecules for untargeted MS-MS spectra. Our approach is based on the premise that molecules and spectra are two different views of the same interaction. This valuable insight allows us to embed molecules and their matching spectra close in the molecule-spectrum joint embedding space. Our approach avoids the need for any kind of reconstruction to intermediate forms such as fingerprints or spectra and therefore removes any reconstruction loss invariably creeping into the ranking pipeline with any reconstruction based approach. The ranking of candidate structures can be attained by comparing their embeddings against that of the query spectrum and selecting the candidate with the highest cosine similarity. The idea of learning joint embedding spaces from multiple views of the data dates back to the seminal work on Siamese Networks [37]. More recently, CLIP (Contrastive Language- Image Pre-training) was trained to create a shared embedding space for both images and text, enabling the model to match relevant images and captions without the need for direct labeling or supervised training on specific dataset [35]. As we embed molecules and spectra in a joint embedding space, our method is termed Joint Embedding Space Technique for Ranking candidate molecules, JESTR. We use CMC, Contrastive Multiview Coding [36], to learn view-invariant information across different views of the data and produce embeddings in a joint embedding space. Recent work in contrastive learning [38, 36, 39] has extended the Siamese network approach significantly. Non-contrastive approaches are also proving to be very effective in learning of such embedding spaces [40, 41, 42, 43]. Broadly speaking, in all these approaches the key question is how to generate the paired views (either appearing naturally or generated via data augmentation); and how to ensure paired views end up close together in the joint embedding space.

Another novelty of our approach lies in utilizing regularization on additional data consisting of millions of molecules with the same chemical formulas as those in the training

dataset. While this form of data augmentation does not contribute directly to additional labeled training data [44], the additional data can be utilized to distinguish target molecules from their candidates. Here, regularization is used as a fine-tuning strategy towards the end of training. When combined with contrastive loss, regularization with additional data provides two key benefits: it improves model generalization by training on a larger, more diverse set of molecules, and it enhances representation learning for both molecule and spectra embeddings by using non-congruent pairs as additional data during training.

For all three problems studied, the main contributions of this thesis are:

- For redox prediction, demonstrating a GPR model trained on the combination of the molecular descriptors, DFT energies, and ADME properties, consistently yields the highest performance on the three test sets, with an R^2 value ranging from 0.84 to 0.94.

- For redox prediction, novel evaluation methodology of using multi-fold cross validation to train and select a generalizable GPR model for redox potential predictions for 14 redox-active quinone molecules gathered by Li et al. [10, 45]. We test these molecules on models trained on benzoquinones and naphthoquinones and report R^2 values ranging from 0.69 to 0.84.

- For redox prediction, novel use and evaluation of the ADME pharmacokinetic properties to predict redox potential.

- For redox prediction, systematic comparison of the effectiveness of molecular properties in predicting redox potential across three datasets using GPR; an ablation study was performed to select the optimal combination of properties.

- In CSI, a generalizable data stratification for view selection on interacting objects, where stratification is applied either on each of the items involved in the interaction in the context of the other object, or on features of the interaction itself.

- Congruent and non-congruent data views allow CSI to be paired with contrastive learning schemes, such as CMC, resulting in learned embeddings suited for downstream tasks.

- Demonstrating how CSI applies to the compound-protein interaction prediction task for protein-drug and to enzyme-compound datasets. The latter dataset is rich in auxiliary interaction information lending itself to stratification on interaction features.

- Showing that CSI significantly outperforms a baseline model not using CSI, where average precision (AP) is improved by 18.2% on the BindingDB dataset, 39% on the BRENDA dataset, and 13.7% on the KEGG dataset, when stratifying by compound and by sequence.

- With JESTR, we present a novel implicit formulation of the annotation problem to avoid explicit prediction of spectra and fingerprints has dominated the field since earliest attempts in solving the problem [46]. Our formulation is grounded in the novel insight that molecules and spectra are views of the same object, similar to recent advances in linking text/image data.

- Demonstrating the effectiveness of contrastive learning in creating a joint embedding space for molecules and spectra, and sufficiency of cosine similarity of the embeddings for ranking the candidate molecules. That is, there is no need for an explicit (learnable) downstream ranking task.

- Fine-tuning the implicit JESTR model via regularization using the candidate sets of the training molecules improves the rank @1 performance in the range of 6.04% to 37.11% when compared to a baseline that does not utilize regularization.

- Demonstrating that JESTR outperforms ESP and MIST on all ranking metrics and all datasets with the exception of rank@1 for the MoNA dataset. For rank@[1 through 5], JESTR outperforms ESP by 71.6% and MIST by 23.6% across three datasets.

These remarkable improvements are achieved even though JESTR does not utilize the additional data in the form of chemical formulae labels for spectra peaks that is currently used by MIST.

1.5 Thesis organization

This thesis investigates methods to create high-quality feature representations to address three important biological problems: redox potential prediction, compound-enzyme interaction prediction, and candidate ranking for metabolite annotation. In chapter 2, we examine prior computational methods developed to address these problems and highlight relevant datasets available for model training. In Chapter 3, we explore how various molecule features contribute to accurate redox using a GPR-based method. In chapter 4, we present CSI, a method utilizing contrastive learning to learn high quality representations for molecules and enzymes and predict the interaction likelihood between them. In chapter 5, we present JESTR, a method to map spectrum and molecule representations into a joint space and rank candidate molecules based on cosine similarity between spectrum and candidates. In chapter 6, we provide a summary of this thesis and outline future directions.

Chapter 2 - Background

The Bioinformatics problems investigated in this thesis vary not just by the nature of the problem but also by the type and amount of available data. In this chapter, we review prior methods developed to address these problems, highlighting available datasets and feature design.

2.1 High-throughput prediction of redox potential

Several methods have been proposed for the computational prediction of redox potentials of organic molecules. As directly related to the molecule's ability to donate or accept electrons, orbital energies have emerged as a key factor in predicting redox potential. The highest occupied molecular orbital (HOMO) energy reflects the ionization energy due to the most loosely held electron in a molecule. Molecules with higher HOMO energies have a lower oxidation potential, and can more easily donate an electron. The LUMO (lowest unoccupied molecular orbital) energy reflects the lowest energy orbital that can accept an electron, and hence electron affinity. Molecules with low-energy LUMOs are more likely to accept an electron because the extra electron is stabilized more effectively. Three different methods are used to calculate orbital energies, trading accuracy for computation time. The Hartree-Fock method can be used to simplify and solve the many-electron Schrodinger equation by treating each electron independently in the presence of all other electrons [47, 48]. This method solves for the wavefunction of each electron, and is computationally expensive.

The Density Functional Theory (DFT) method works on the principle that HOMO/LUMO energies can be determined by estimating electron density, rather than the wavefunctions [49]. This method scales better in performance as compared to the theoretical Hartree-Fock method because of the simplifying assumptions. Semi-empirical methods are quantum mechanical methods where some integrals are simplified based on experimental data to reduce computational cost by sacrificing accuracy [50, 51]. All these methods essentially solve functionals for all the electron-electron interactions, and for all the orbitals in the molecule. As such, these methods do not scale well for large molecular systems and hence are not suitable for high-throughput screening of large compound libraries.

High throughput redox potential predictions require data driven approaches [52] to train models on molecular features. There are a number of relevant features which can be used to predict redox potential, such as Morgan fingerprints [53], molecular descriptors [54, 55] and molecular orbital energies [55]. Prior work [56] introduced the concept of combining features of the neutral molecules with those of the reduced molecule to form a reaction fingerprint, which can then be used as an input to the model. Such reaction fingerprints are known to be better at capturing reaction features compared to individual molecular features [56].

One challenge with redox potential prediction models is the paucity of ground truth experimental data required to train the model. The number of experimental values available for redox potentials is usually on the order of a few hundred. As a result, deep learning models, which work best with large training datasets, are difficult to effectively train for redox prediction. In contrast, simple models such as a linear fit would likely overfit on such small datasets. Therefore, an intermediate-sized model is needed to address both of these issues. Gaussian Process Regression (GPR) is a widely used method for regression and classification tasks [57, 55, 23, 58] since it can generate a distribution of functions which fit the observed data [59]. GPR models treat each data point as part of a multivariate distribution, enabling accurate predictions on both small and large datasets with improved

resiliency to outliers.

2.2 Compound protein interaction prediction

State-of-art machine-learning methods for drug-target interaction likelihood prediction have been extensively reviewed in recent survey papers, e.g., [60, 61, 62, 63, 64, 65]. Predicting the interaction between a compound and protein sequence elucidates drug-protein interactions [61] and promiscuous enzymatic activities on substrates [66]. Related deep-learning methods broadly perform two tasks: representation learning of compounds and of protein sequences, and using the learned representations to predict interactions. Molecular representations can be learned from molecular fingerprints [67, 68, 69] or learned on the corresponding molecular graphs using Graph Neural Networks (GNNs) [70, 71]. Deep learning models such as CNNs [69], and transformers [72, 73] are used to generate embeddings on protein sequences. Recent models also incorporate 3-D structure of the enzymes to make the representations better [74]. Interaction models however remain simple, where representations are concatenated, with or without attention, to predict interaction likelihood. Cross attention between molecule features and enzyme features have also been used to learn better interaction models [75].

Unlike 3-D docking simulations [76], deep-learning models allow screening a large number of putative interactions efficiently. While these models attempt to learn better embeddings for the molecule and the enzyme, the power of modeling different views of the interactions is not exploited. Our CSI model focuses on utilizing different views of interaction to overcome this limitation. The next challenge is selecting the relevant datasets for training and testing our model. We train and evaluate CSI models for three datasets. The BindingDB dataset [77] contains purchasable drugs and their protein targets which exhibit an affinity higher than 10 μM, and is larger and more diverse than earlier drug-protein interaction datasets. The BRENDA dataset is derived from the BRENDA database [78], which provides continued manual and automated curation on enzymes and compounds

interacting with enzymes. The KEGG dataset is derived from the KEGG database [79], which catalogues biochemical reactions for a large set of organisms. Other datasets such as KIBA [80] and Davis [81] have been used in many prior works, but are limited in the molecules and enzymes they cover, and have also not been updated in recent times.

2.3 Candidate ranking for untargeted metabolomics spectra

Metabolomics workflows are complex and require multiple prepossessing steps on the raw data files collected from the mass spectrometry instruments. For example, MZmine [82] offers a wealth of tools for raw data smoothing and filtering, peak detection, alignment of various measured features, and many others to generate the spectrum data. There are now tools for *de novo* molecular formula determination of a query spectrum [83] that narrows the molecular formula search space compared to prior approaches [84], yielding 93.0% accuracy for m/z values less than 400 Da, and 63.8% accuracy for m/z $\geq$ 400 Da. The measured mass spectrum is a collection of peaks (Figure 2.1A). Each peak is represented by its mass-to-charge (m/z) ratio, where the charge is known and is often +1 or -1, and a relative intensity. Several techniques address the problem of annotating spectra. The go-to technique is "spec-to-spec" comparison (Figure 2.1B) of the measured query spectra against spectra which are catalogued in spectral reference libraries [85]. However, despite the growth in spectral libraries, e.g., GNPS [86], NIST [87], MoNA [88], annotation rates remain extremely low due to the imitated coverage of spectral libraries in comparison to the space of all potential molecules. In addition, measured spectra vary tremendously under differing instrument settings, e.g., ionization energy, solvent type, and adduct formation (additional functional groups attached or removed from the ionized molecule). A molecule therefore may have many corresponding spectra, which further limits library coverage. A recent search of spectra within 15,327 datasets [86] deposited in the MassIVE (Mass Spectrometry Interactive Virtual Environment) database against 586,647 reference spectra catalogued in the GNPS reference library yielded a positive identification rate of 2.3% [89]

(Figure 2.1C).

Two types of supervised predictive annotation techniques have emerged. "Mol-to-spec" techniques (Figure 2.1D) utilize combinatorial fragmentation approaches, e.g., MetFrag [90, 91] and CFM-ID [92], MLPs [93], or GNNs [94, 95, 96], to translate a molecular structure into a predicted spectrum. Candidate molecular structures are retrieved by either chemical formula, if available, or molecular mass from large molecular databases such as PubChem [97], or more biologically relevant, smaller, databases. The candidate with the most similar spectrum to the query spectrum is ranked highest and used as the annotation. In contrast, "spec-to-mol" techniques (Figure 2.1E) aim to generate *de novo* molecular candidates which potentially match the query spectrum, e.g., MSNovelist [98], Spec2Mol [99], MS2Mol [100]. For example, MS2Mol uses sequence-to-sequence transformers to translate spectra into *de novo* molecular structures in the form of SMILES strings (a description of the chemical structure using an ASCII strings). Due to their current limited capabilities, *de novo* generation is currently of limited use in the metabolomics community. An alternative and earlier approach is "spec-to-FP" (Figure 2.1F), where a molecular fingerprint (FP) vector is predicted for the query spectrum, e.g., Sirius [84], MIST [101]. Here, the predicted fingerprint is compared against those of the candidate molecular structures, and the best match, via Tanimoto or cosine similarity, is declared the annotation result. Despite recent advances in all such techniques, annotation rates remain low as the *reconstruction* of spectrum, fingerprint, or molecular structure is a difficult task.

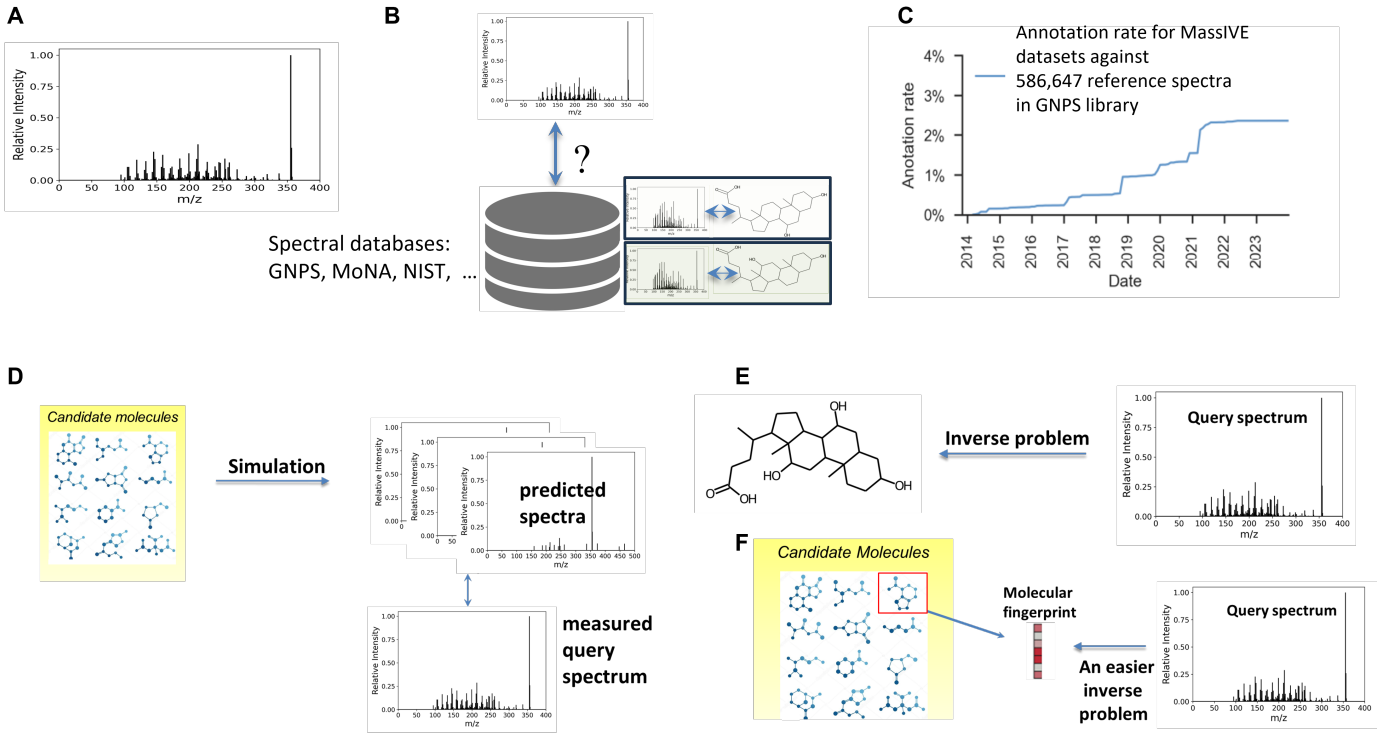


Figure 2.1 Current annotation workflows. A. Example spectrum measured using LC-MS or MS/MS, where x-axis represent the mass-to-charge ratio and the y-axis represent the relative intensity of each peak. B. Reference library search using spec-to-spec comparison. C. Current annotation rates using state-of-the-art library search are low despite growth in reference databases. D. Mol-to-spec predictive approach mimics the mass spectrometry fragmentation process. E. Spec-to-mol involves *de novo* molecular generation from spectra and forms an inverse problem. F. Spec-to-FP approach predicts a molecular fingerprint and identifies the candidate structure that most likely matches the predicted fingerprint.

Chapter 3 - The role of structural, pharmacokinetic and energy properties in the high-throughput prediction of redox potentials for organic molecules with experimental calibration

Reduction/Oxidation (redox) potential of small organic compounds is a key property that drives innumerable chemical and biological electron transfer reactions. However, experimental measurement of redox potential is time-consuming and expensive, yielding few and small experimental measured datasets. Computational methods have previously been applied to create redox predictors applicable only to a specific dataset. We investigate the effectiveness of various descriptors, including structural and functional properties, molecular energies, and drug-like properties, to predict redox potential. We use Gaussian Process Regression (GPR) as a model, as it is suitable for fitting small datasets. We train and test our redox predictor on three organic molecule datasets. We demonstrate that a GPR-based redox predictor using a combination of molecular descriptors, DFT energies, and pharmacokinetic properties works well across the datasets. Finally, we test the trained model against an experimental dataset of quinones to show that the model makes predictions well correlated with experimental data.

3.1 Methods

An overview of our method is provided in Figure 3.1.

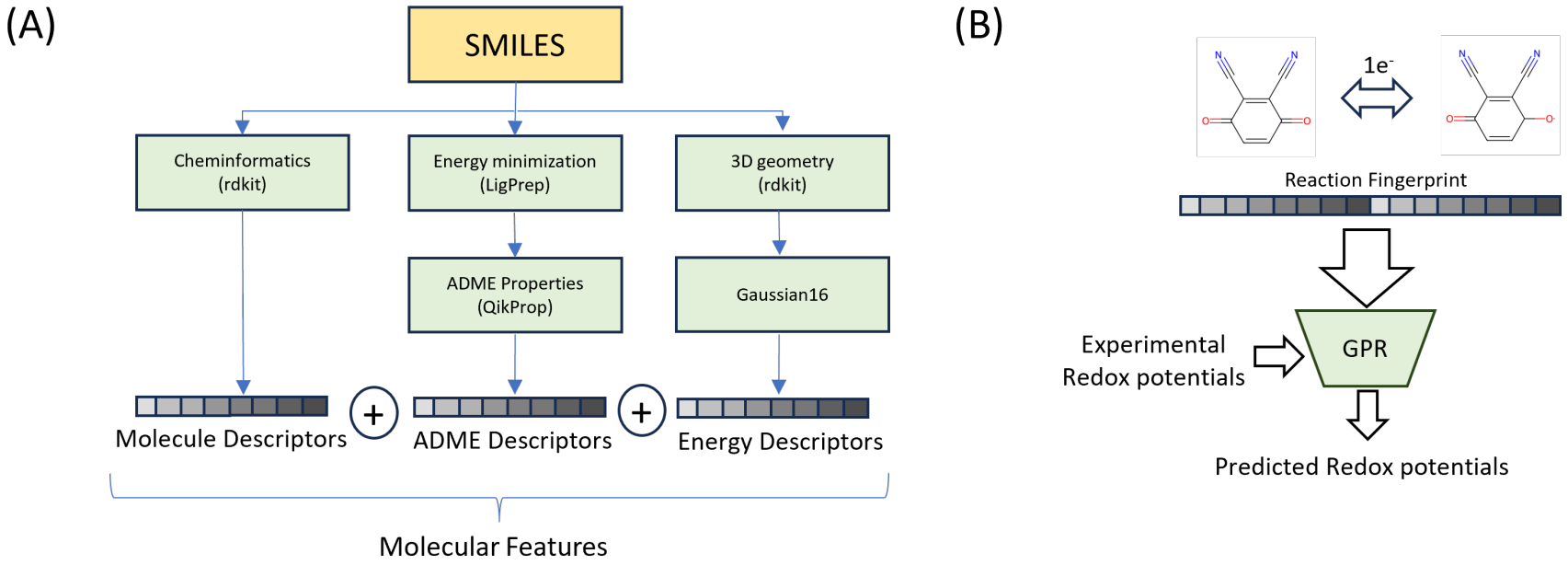


Figure 3.1 Method overview. (A) The molecular features are computed using various methods. One or more molecular features are concatenated into a single vector that represents each molecule. (B) The GPR model is trained to predict redox potential based on the reaction fingerprint as an input. The model is trained against the experimental values of redox potential as ground truth.

3.1.1 Molecular features

Molecular features are vector representations that encode structural and functional properties of molecules. We investigate different types of molecular fingerprints for their impact on redox potential prediction, both individually and in combination.

Morgan fingerprints

Extended Connectivity fingerprints (ECFP) are a class of topological fingerprints for molecule characterization [53]. ECFP are circular fingerprints whose features represent the presence of specific substructures in the molecule. The fingerprints are calculated using the Morgan algorithm [102], hence ECFP are also called Morgan fingerprints. Substructures are determined by exploring the neighbors of atoms in a molecule iteratively. The user can control the radius of these iterative circles to gather larger substructures. Each substructure is assigned a numerical identifier and all identifiers are accumulated into an array and then hashed into a unique number of specific bit size. For this work, we used a radius of 4 as an optimal balance between including sufficient connections from each atom and computational complexity. We use 512-bit fingerprints for our experiments.

Molecular descriptors

Molecular descriptors are features that include chemical properties, such as chirality information, charge related information, bond type and cycle information, in addition to structural information. The cheminformatics module within the rdkit package [54] calculates a set of 208 such molecular descriptors using the Simplified Molecular Input Line Entry System (SMILES) [103] string of the molecule as input. The 208-dimensional descriptors were scaled using MinMaxScaler in scikit-learn [104] to fit the values within the range of 0 and 1. As described in Ghule et al [55], the list of all molecular descriptors obtained from rdkit [54] is given below:

MaxEStateIndex, MinEStateIndex, MaxAbsEStateIndex, MinAbsEStateIndex, qed,

MolWt, HeavyAtomMolWt, ExactMolWt, NumValenceElectrons, NumRadicalElectrons, MaxPartialCharge, MinPartialCharge, MaxAbsPartialCharge, MinAbsPartialCharge, FpDensityMorgan1, FpDensityMorgan2, FpDensityMorgan3, BCUT2D_MWHI, BCUT2D_MWLOW, BCUT2D_CHGHI, BCUT2D_CHGLO, BCUT2D_LOGPHI, BCUT2D_LOGPLOW, BCUT2D_MRHI, BCUT2D_MRLOW, BalabanJ, BertzCT, Chi0, Chi0n, Chi0v, Chi1, Chi1n, Chi1v, Chi2n, Chi2v, Chi3n, Chi3v, Chi4n, Chi4v, HallKierAlpha, Ipc, Kappa1, Kappa2, Kappa3, LabuteASA, PEOE_VSA1, PEOE_VSA10, PEOE_VSA11, PEOE_VSA12, PEOE_VSA13, PEOE_VSA14, PEOE_VSA2, PEOE_VSA3, PEOE_VSA4, PEOE_VSA5, PEOE_VSA6, PEOE_VSA7, PEOE_VSA8, PEOE_VSA9, SMR_VSA1, SMR_VSA10, SMR_VSA2, SMR_VSA3, SMR_VSA4, SMR_VSA5, SMR_VSA6, SMR_VSA7, SMR_VSA8, SMR_VSA9, SlogP_VSA1, SlogP_VSA10, SlogP_VSA11, SlogP_VSA12, SlogP_VSA2, SlogP_VSA3, SlogP_VSA4, SlogP_VSA5, SlogP_VSA6, SlogP_VSA7, SlogP_VSA8, SlogP_VSA9, TPSA, EState_VSA1, EState_VSA10, EState_VSA11, EState_VSA2, EState_VSA3, EState_VSA4, EState_VSA5, EState_VSA6, EState_VSA7, EState_VSA8, EState_VSA9, VSA_EState1, VSA_EState10, VSA_EState2, VSA_EState3, VSA_EState4, VSA_EState5, VSA_EState6, VSA_EState7, VSA_EState8, VSA_EState9, FractionCSP3, HeavyAtomCount, NHOHCount, NOCount, NumAliphaticCarbocycles, NumAliphaticHeterocycles, NumAliphaticRings, NumAromaticCarbocycles, NumAromaticHeterocycles, NumAromaticRings, NumHAcceptors, NumHDonors, NumHeteroatoms, NumRotatableBonds, NumSaturatedCarbocycles, NumSaturatedHeterocycles, NumSaturatedRings, RingCount, MolLogP, MolMR, fr_Al_COO, fr_Al_OH, fr_Al_OH_noTert, fr_ArN, fr_Ar_COO, fr_Ar_N, fr_Ar_NH, fr_Ar_OH, fr_COO, fr_COO2, fr_C_O, fr_C_O_noCOO, fr_C_S, fr_HOCCN, fr_Imine, fr_NH0, fr_NH1, fr_NH2, fr_N_O, fr_Ndealkylation1, fr_Ndealkylation2, fr_Nhpyrrole, fr_SH, fr_aldehyde, fr_alkyl_carbamate, fr_alkyl_halide, fr_allylic_oxid, fr_amide, fr_amidine, fr_aniline, fr_aryl_methyl, fr_azide, fr_azo, fr_barbitur, fr_benzene, fr_benzodiazepine, fr_bicyclic, fr_diazo, fr_dihydropyridine, fr_epoxide, fr_ester, fr_ether, fr_furan, fr_guanido, fr_halogen, fr_hdrzine, fr_hdrzone, fr_imidazole, fr_imide, fr_isocyan, fr_isothiocyan, fr_ketone, fr_ketone_Topliss, fr_lactam, fr_lactone, fr_methoxy, fr_morpholine, fr_nitrile, fr_nitro, fr_nitro_arom,

fr_nitro_arom_nonortho, fr_nitroso, fr_oxazole, fr_oxime, fr_para_hydroxylation, fr_phenol, fr_phenol_noOrthoHbond, fr_phos_acid, fr_phos_ester, fr_piperdine, fr_piperzine, fr_priamide, fr_prisulfonamd, fr_pyridine, fr_quatN, fr_sulfide, fr_sulfonamd, fr_sulfone, fr_term_acetylene, fr_tetrazole, fr_thiazole, fr_thiocyan, fr_thiophene, fr_unbrch_alkane, fr_urea

DFT energies

Orbital energies are closely correlated to redox potentials, and have been used in prior works as one of the features to predict redox potential [105, 57, 6, 55]. Therefore, despite the high computational cost, we explored orbital energies calculated using DFT methods as one of the features. Quinones undergo two distinct single-electron reductions, and two protonation steps to yield the fully reduced hydroquinones. From neutral quinones, the radical anion is the addition of one electron (1e-), the di-anion is the addition of two electrons (2e-), while the protonated anion is the addition of two electrons and one proton (2e-, 1H), and addition of two electron and two hydrogens (2e-, 2H) ultimately results in fully reduced hydroquinone [106]. As a result, how stability electrons can be housed during reduction steps (HOMO and LUMO energies) is an important contributor to a compound's redox behavior. Indeed, orbital energies calculated with DFT approaches have been shown to be effective predictors for redox potential across a wide spectrum of chemical species, including π-conjugated molecules [52]. Energy based descriptors are therefore applicable to organic compounds with π-bonds. We used the Gaussian16 package with B3LYP hybrid functional [107] in combination with the 6-311+G(2d,p) basis set to calculate the HOMO/LUMO energies as a single point calculation. The starting geometries for the neutral compounds were calculated using the distance geometry package in rdkit. The energy values were scaled using the MinMaxScaler.

ADME properties

ADME properties include pharmacokinetic and physicochemical properties such as octanol/water and water/gas logP, logS, logBB, overall central nervous system (CNS) activity, Caco-2 and MDCK cell permeabilities, logKhsa for human serum albumin binding, and log IC50 for HERG K+-channel blockage. Such properties are known to play an important role in drug metabolism [108]. Drugs with suitable redox potentials are efficiently oxidized or reduced by enzymes, such as cytochrome P450 thereby indicating a correlation between ADME properties and redox potential [108]. The mechanism of action for Mitoxantrone, Daunorubicin and Tenofovir involves redox chemistry and this plays a crucial role in their binding properties with specific enzymes. ADME properties are a good measure of drug-enzyme binding affinity [109]. We predicted 52 ADME properties using the Schrodinger QikProp tool [24]. To prepare molecules for QikProp processing, we used the Schrodinger LigPrep utility, which creates three-dimensional ligand structures and performs molecular mechanics minimization [110]. The 3D structures are then input into the QikProp tool [24] which predicts physically significant descriptors and pharmaceutically relevant properties. For certain molecules, LigPrep may generate multiple tautomers and/or stereoisomers: in those cases, we applied QikProp to every structure corresponding to the molecule and took the average value for each individual property. The ADME properties were scaled using the MinMaxScaler. The ADME properties generated using QikProp [24] are given below:

Molecule_name, #stars, #amine, #amidine, #acid, #rotor, #rtvFG, CNS, mol_MW, dipole, SASA, FOSA, FISA, PISA, WPSA, volume, donorHB, accptHB, dip2/V, ACxDN0.5/SA, glob, QPpolrz, QPlogPC16, QPlogPoct, QPlogPw, QPlogPo/w, QPlogS, QPlogHERG, QPPCaco, QPlogBB, QPPMDCK, QPlogKp, IP(ev), EA(eV), #metab, QPlogKhsa, HumanOralAbsorption, PercentHumanOralAbsorption, SAFluorine, SAamideO, PSA, #NandO, RuleOfFive, RuleOfThree, #ringatoms, #in34, #in56, #noncon, #nonHatm, Jm

3.1.2 Reaction fingerprints

Each redox reaction can be modeled as a transformation that converts the original molecule into its reduced form. To create a vector representation of a reaction, and following the work of [58, 56], we concatenate the molecular features of the original molecule and those of the reduced molecule (Figure 3.1A). Reaction fingerprints are used as an input to our GPR predictor.

3.1.3 GPR model

We train a GPR model [59] for our predictions (Figure 3.1B). GPR is a probabilistic supervised machine learning model. It uses prior knowledge in the form of kernels and can provide uncertainty measures over predictions. Given any data, we can theoretically fit an infinite set of functions to that data. GPR trains by defining a distribution of parameters over this exhaustive list of functions. Each distribution is modeled as a multivariate normal distribution, and the parameters of this distribution are learned during the training process. GPR models are shown to be more resilient to outliers and fit better on small datasets [111]. We used GPR from the sklearn package with the Radial Basis Function (RBF) kernel as the core kernel as well as White Kernel to model noise [112]. The RBF kernel is defined as:

$$k(x_i, x_j) = exp(-\frac{d(x_i, x_j)^2}{2l^2}) \tag{3.1}$$

where x_i and x_j are the inputs and l is the length scale of the kernel varied between limits given by length_scale_bounds. In our case, we used length_scale = 1.0 and length_scale_bounds = (10^{-3}, 10^3).

The gaussian noise is inserted using a WhiteKernel, which is defined as

$$k(x_i, x_j) = noise\_level\ if\ x_i = x_j\ else\ 0 \tag{3.2}$$

For the WhiteKernel, we used noise_level = 1.0 and noise_level_bounds = (10^{-10}, 10^{1})

The overall kernel for the GPR is then defined as

$$kernel = 1.0 * RBF + WhiteKernel \tag{3.3}$$

We used various molecular features individually and in combination to create the reaction fingerprints. The reaction fingerprints were then used to train the GPR model using each of the three datasets as a training set. 10% of the data was randomly kept aside as the test set. We used a 5-fold random split protocol to help evaluate our model. This approach splits the dataset into 5 consecutive random folds of train and test. We report average metrics over the 5 folds. For testing the experimental set, the entire dataset is used for testing. R^2 (coefficient of determination) was used as the metric to measure the accuracy of predictions against the experimentally measured values of redox potentials.

3.2 Experiments and results

3.2.1 Datasets

We used three datasets to evaluate our GPR-based model. The first two datasets comprise quinones (Figure 3.2) while the third dataset comprises organic compounds with other redox active functionalities. Quinones are prevalent in biological reactions with both 1,4-benzoquinones (e.g. ubiquinone) and 1,4-naphthoquinones (e.g. menaquinone) playing critical biological roles. Both ubiquinone and menaquinone are endogenous cofactors for proteins involved in the electron transport chain and act as antioxidants to protect cell membranes. Both contain two alkyl substitutions, a methyl and a long prenyl chain. The latter significantly increases the lipophilicity of these cofactors and this sequesters them into the cell membranes. Beyond cofactors, hundreds of other quinones have been isolated from natural sources and these molecules are known to play an important role in redox chemistry for both prokaryotes and eukaryotes. The following libraries represent the

structural diversity found in nature along with general redox active organic compounds to test the generalizability of our method.

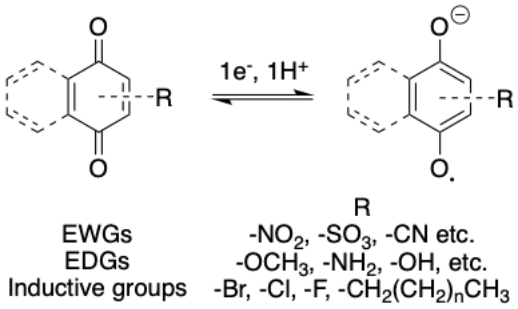


Figure 3.2 Representation of benzoquinone and naphthoquinone datasets and the 1e- reduction

Benzoquinone dataset

The single electron redox potentials of 104 structurally diverse benzoquinones (Table 3.1) were previously measured by Prince et al. using cyclic voltammetry in dimethylformamide (DMF) [22]. This library consist of primarily molecules with molecular weights 122.1 - 611.7 Da and incorporates diverse functionalities that are conjugated to the quinone moiety, including electron withdrawing groups (e.g., cyano, nitro, aldehyde, haloalkyl), electron donating groups (e.g., hydroxy, alkoxy, amine), and inductive groups (e.g., halogens, alkyl), thus yielding large ranges of measured redox potentials (-1077 mV to 597 mV). We removed five compounds that contain iodine, since the presence of iodine complicates the generation of DFT energy based features.

Naphthoquinone dataset

The naphthoquinone dataset [22] used the same setup as above to measure single electron redox potentials for 79 naphthoquinones (Table 3.2). Overall, this library represents compounds with molecular weight 172.2 - 488.7 Da and the structural features of this library can be split into two sub-groups, those in the 2,3-position that are directly conjugated to the quinone moiety and those on the aromatic ring. Similar to the benzoquinone library, the

Table 3.1 Benzoquinone molecule SMILES in the dataset

O=C1C=CC(=O)C(=C1)C#N	CC(C)C1=C(C(=O)C(=C(C1=O)[N+]([O-])=O)C)[N+]([O-])=O	CC1=C(C(=O)C(=C(C1=O)[S](=O)(=O)c2ccccc2)C)[S](=O)(=O)c3ccccc3
O=C1C=CC(=O)C(=C1C#N)C#N	CC1=C(C(=O)C=CC1=O)C(C)(C)c2ccccc2	CC1=C(C(=O)C(=C(C)C1=O)[S](=O)(=O)c2ccccc2)[S](=O)(=O)c3ccccc3
ClC1=C(Cl)C(=O)C(=C(C#N)C1=O)C#N	CC1=C(C)C(=O)C=CC1=O	FC(F)(F)C1=CC(=O)C=CC1=O
ClC1=CC(=O)C=CC1=O	CC1=CC(=O)C(=CC1=O)C	O=C1C=CC(=O)C(=C1)C(=O)c2ccccc2
CC(C)(C)C1=CC(=O)C(=CC1=O)Br	CC1=CC(=O)C=C(C)C1=O	CC1=C(C(=O)C=CC1=O)C(=O)c2ccccc2
CC1=CC(=O)C=C(Br)C1=O	CC(C)=CCC1=CC(=O)C(=C(C)C1=O)C	CC1=CC(=O)C(=CC1=O)C(=O)c2ccccc2
CC(C)C1=CC(=O)C=C(I)C1=O	CC(C)C1=C(C(C)C)C(=O)C(=C(C)C1=O)C	COC1=CC(=O)C=CC1=O
CC(C)C1=CC(=O)C(=C(I)C1=O)C	CC1=C(C)C(=O)C2=C(OCO2)C1=O	COC1=C(C)C(=O)C=CC1=O
ClC1=CC(=O)C=C(Br)C1=O	CC(C)(C)C1=CC(=O)C(=CC1=O)C(C)(C)C	COC1=C(OC)C(=O)C=CC1=O
ClC1=C(Cl)C(=O)C=CC1=O	CC(C)(C)C1=CC(=O)C=C(C1=O)C(C)(C)C	COC1=CC(=O)C(=CC1=O)OC
ClC1=CC(=O)C(=CC1=O)Cl	CC(C)C1=CC(=O)C=C(C(C)C)C1=O	COC1=CC(=O)C=C(OC)C1=O
ClC1=CC(=O)C=C(Cl)C1=O	CC(C)CCCCCC1=CC(=O)C(=CC1=O)CCCCCC(C)C	COC1=C(OC)C(=O)C(=CC1=O)C
BrC1=CC(=O)C=C(Br)C1=O	CC1=CC(=O)C(=C(C)C1=O)C	COC1=C(OC)C(=O)C(=C(Br)C1=O)Br
IC1=CC(=O)C=C(I)C1=O	CC(C)CCC[C@@H](C)CCC[C@@H](C)CCCC(/C)=C/CC1=C(C)C(=O)C(=C(C)C1=O)C	COC1=C(OC)C(=O)C(=C(C)C1=O)C
CC(C)(C)C1=CC(=O)C(=C(Cl)C1=O)Cl	CCC(C)(C)C(O)=O.O=C1C=CC(=O)C=C1	CCCCCCCCCCC1=C(C)C(=O)C(=C(OC)C1=O)OC
CC(C)(C)C1=CC(=O)C(=C(Br)C1=O)Br	CC1=C(C)C(=O)C(=C(C)C1=O)SCCC(O)=O	COC1=C(OC)C(=O)C(=C(C)C1=O)CC=C(C)C
CC(C)(C)C1=CC(=O)C(=C(I)C1=O)I	CC1=C(C)C(=O)C(=C(C)C1=O)C	CC(O)C1=CC(=O)C=CC1=O
CC1=C(Cl)C(=O)C(=C(C)C1=O)Cl	CC(C)C1=C(C(C)C)C(=O)C(=C(C(C)C)C1=O)C(C)C	COC(=O)C1=CC(=O)C=CC1=O
CC(C)(C)C1=C(Cl)C(=O)C(=C(Cl)C1=O)C(C)(C)C	CC(=O)C1=CC(=O)C=CC1=O	COC(=O)C1=C(C(=O)C=CC1=O)[S](=O)(=O)c2ccccc2
CC(C)C1=C(Br)C(=O)C(=C(Br)C1=O)C	CC(=O)C1=C(C)C(=O)C=CC1=O	COC(=O)C1=CC(=O)C(=CC1=O)C(=O)OC
CC(C)C1=C(Br)C(=O)C(=C(Br)C1=O)C(C)C	CC(=O)C1=CC(=O)C(=CC1=O)C	CCNC1=CC(=O)C(=CC1=O)NCC
BrC1=C(C(=O)C(=C(C1=O)c2ccccc2)Br)c3ccccc3	CC(=O)C1=CC(=O)C=C(C)C1=O	O=C1C=CC(=S)C(=C1)c2ccccc2
COC1=C(Cl)C(=O)C(=C(Cl)C1=O)OC	CC(=O)C1=C(Cl)C(=O)C=CC1=O	CC(C)SC1=CC(=O)C(=CC1=O)C(C)(C)C
COC1=C(Br)C(=O)C(=C(OC)C1=O)Br	CC(C)(C)C(=O)C1=CC(=O)C=CC1=O	CC(C)NC1=CC(=O)C=C(C1=O)C(C)(C)C
NC1=C(Br)C(=O)C(=C(Br)C1=O)N	O=C1C=CC(=O)C(=C1)c2ccccc2	CC(C)(C)NC1=CC(=O)C(=CC1=O)C(C)(C)C
BrC1=CC(=O)C(=CC1=O)Br	O=C1C=CC(=O)C(=C1c2ccccc2)c3ccccc3	CC(C)(C)NC1=CC(=O)C=C(C1=O)C(C)(C)C
ClC1=CC(=O)C(=C(Cl)C1=O)Cl	O=C1C=C(C(=O)C=C1c2ccccc2)c3ccccc3	CC(C)(C)NC1=CC(=O)C(=CC1=O)NC(C)(C)C
CCCCCCC1=C(Br)C(=O)C(=C(Br)C1=O)Br	O=C1C=CC(=O)C(=C1)COc2ccccc2	CC1=C(O)C(=O)C=CC1=O
FC1=C(F)C(=O)C(=C(F)C1=O)F	O=C1C=CC(=O)C(=C1)[S](=O)(=O)c2ccccc2	CC1=C(O)C(=O)C(=C(Br)C1=O)Br
ClC1=C(Cl)C(=O)C(=C(Cl)C1=O)Cl	CC(=O)C1=C(C(=O)C=CC1=O)[S](=O)(=O)c2ccccc2	CCCCCCCCCCCCC1=C(C)C(=O)C(=C(C)C1=O)O
BrC1=C(Br)C(=O)C(=C(Br)C1=O)Br	CC(C)(C)C(=O)C1=C(C(=O)C=CC1=O)[S](=O)(=O)c2ccccc2	OC1=CC(=O)C(=CC1=O)O
IC1=C(I)C(=O)C(=C(I)C1=O)I	O=C1C=CC(=O)C(=C1C(=O)c2ccccc2)[S](=O)(=O)c3ccccc3	OC1=C(Cl)C(=O)C(=C(Cl)C1=O)O
CC1=CC(=O)C=CC1=O	CC1=C(C)C(=O)C(=CC1=O)[S](=O)(=O)c2ccccc2	O=C1C=C(N2CCCCC2)C(=O)C=C1N3CCCCC3
CC(C)(C)C1=CC(=O)C=CC1=O	CC1=C(C)C(=O)C(=C(C)C1=O)[S](=O)(=O)c2ccccc2	COC(=O)C1=C(Cl)C(=O)C=CC1=O
CC(C)C1=CC(=O)C(=CC1=O)C	O=C1C=C(C(=O)C=C1[S](=O)(=O)c2ccccc2)[S](=O)(=O)c3ccccc3	

naphthoquinone library consists of a large diversity of functionality at the 2,3-position of the quinone, including EWGs (e.g., sulfonates), EDGs (e.g., hydroxy, methoxy), and inductive groups (e.g., halogens, alkyl). The library has less functionalization of the aromatic ring, with incorporation of EDG (e.g., hydroxyl, methoxy, methyl, cyano groups) and inductive groups (e.g., chloro and bromo groups).

General organic compound dataset

The ROP313 dataset [105] is a widely used dataset with experimentally measured redox potentials for 313 organic (183) and organometallic (130) compounds (Table 3.3). The molecular weights of the organic compounds range from 68.07 - 647.9 Da. One electron redox potential was measured using cyclic voltammetry in acetonitrile (ACN). For this work, we used the organic compounds as our third dataset. We removed eight, four that contained iodine, since the presence of iodine complicates the generation of DFT energy based features, and four compounds that did not have any π-bonds.

The Li test dataset

To further evaluate our model against new data, we compare the predictions against experimental data. We used experimental redox potentials of 14 redox-active quinone molecules (Figure 3.3) measured by Li et al. [10]. Their work investigates quinones involved in the extracellular electron transfer (EET) in certain microorganisms. The authors created a model to predict the EET using redox potential as one of the inputs to the model. The dataset from their paper contains both 4 benzoquinones, 7 naphthoquinones, and 3 anthraquinones with a diverse set of functional groups attached to these quinones. Single electron redox potential was measured using cyclic voltammetry in a single chamber electrochemical cell with carbon and platinum mesh electrodes with Ag/Ag+ as the reference electrode [10].

Table 3.2 Naphthoquinone molecule SMILES in the dataset

O=C1C=CC(=O)c2cc(ccc12)C#N	COC1=CC(=O)c2ccccc2C1=O	CCCCCCCCC1=C(O)C(=O)c2ccccc2C1=O
CC1=C(C)C(=O)c2cc(ccc2C1=O)C#N	COc1cccc2C(=O)C=CC(=O)c12	CCCCCCCCCCC1=C(O)C(=O)c2ccccc2C1=O
ClC1=CC(=O)c2ccccc2C1=O	COC1=CC(=O)c2ccc(O)cc2C1=O	CCCCCCCCCCCCC1=C(O)C(=O)c2ccccc2C1=O
CCCNC1=C(Cl)C(=O)c2ccccc2C1=O	COC1=C(OC)C(=O)c2ccccc2C1=O	CC(C)CCC1=C(O)C(=O)c2ccccc2C1=O
ClC1=C(Cl)C(=O)c2ccccc2C1=O	COc1ccc(OC)c2C(=O)C=CC(=O)c12	CC(C)=CCC1=C(O)C(=O)c2ccccc2C1=O
Clc1cc2C(=O)C=CC(=O)c2cc1Cl	CC(=O)Oc1cccc2C(=O)C=CC(=O)c12	CC(C)CCCC(C)CCC1=C(O)C(=O)c2ccccc2C1=O
BrC1=C(Br)C(=O)c2ccccc2C1=O	CC(=O)Oc1ccc(O)c2C(=O)C=CC(=O)c12	CC(CCC1=C(O)C(=O)c2ccccc2C1=O)c3ccc(Oc4ccccc4)cc3
CC1=CC(=O)c2cc(Br)c(Br)cc2C1=O	CC(=O)Oc1ccc(OC(C)=O)c2C(=O)C=CC(=O)c12	OC1=C(CCCCc2ccc(Oc3ccccc3)cc2)C(=O)c4ccccc4C1=O
CC1=C(C)C(=O)c2cc(Br)c(Br)cc2C1=O	CC(=O)Oc1ccc(OC(C)=O)c2C(=O)C(=CC(=O)c12)OC(C)=O	OC1=C(CCCCCCCCCCBr)C(=O)c2ccccc2C1=O
CC1=CC(=O)c2ccccc2C1=O	C=CCOC1=CC(=O)c2ccccc2C1=O	OC1=C([C@@H]2CC[C@@H](CC2)C3CCCCC3)C(=O)c4ccccc4C1=O
Cc1cccc2C(=O)C=CC(=O)c12	CC(C)OC1=CC(=O)c2ccccc2C1=O	Cc1ccc(SC2=C(O)C(=O)c3ccccc3C2=O)cc1
Cc1ccc2C(=O)C=CC(=O)c2c1	O=C1C=C(C(=O)c2ccccc12)[S](=O)(=O)c3ccccc3	CCCCCCCCCCSCCCCCCCCCCC1=C(O)C(=O)c2ccccc2C1=O
CC(C)(C)C1=CC(=O)c2ccccc2C1=O	NC1=CC(=O)c2ccccc2C1=O	Cc1ccc(SCCCCCCCCCCC2=C(O)C(=O)c3ccccc3C2=O)cc1
O=C1C=C(C2CCCCC2)C(=O)c3ccccc13	CC(=O)NC1=C(N)C(=O)c2ccccc2C1=O	OC1=C(CCCCCCCCCCOc2ccc(cc2)C3CCCCC3)C(=O)c4ccccc4C1=O
CC1=C(C)C(=O)c2ccccc2C1=O	CCC(=O)NC1=C(N)C(=O)c2ccccc2C1=O	OC1(N=NC2=CC(=O)c3ccccc3C2=O)C(c4ccccc4)C(=O)c5ccccc5C1=O
Cc1ccc2C(=O)C(=CC(=O)c2c1)C	CN(C)C1=C(N)C(=O)c2ccccc2C1=O	COc1ccc2C(=O)C=C(O)C(=O)c2c1
Cc1ccc2C(=O)C=C(C)C(=O)c2c1	CN(C)C1=CC(=O)c2ccccc2C1=O	COc1ccc2C(=O)C(=C(O)C(=O)c2c1)CCCC3CCCCC3
Cc1cccc2C(=O)C=C(C)C(=O)c12	CC1=CC(=O)c2ccccc2C1=S	Oc1cccc2C(=O)C=C(Br)C(=O)c12
Cc1ccc(C)c2C(=O)C=CC(=O)c12	CSC1=C(C)C(=O)c2ccccc2C1=O	CC1=CC(=O)c2c(O)cccc2C1=O
Cc1cc2C(=O)C=CC(=O)c2cc1C	CCSC1=C(C)C(=O)c2ccccc2C1=O	CC1=CC(=O)c2cccc(O)c2C1=O
CCC1=C(C)C(=O)c2ccccc2C1=O	OC1=CC(=O)c2ccccc2C1=O	OC1=C(O)C(=O)c2ccccc2C1=O
CCCCCC1=C(C)C(=O)c2ccccc2C1=O	Oc1cccc2C(=O)C=CC(=O)c12	Oc1ccc(O)c2C(=O)C=CC(=O)c12
CCCCCCCCCCC1=C(C)C(=O)c2ccccc2C1=O	CC1=C(O)C(=O)c2ccccc2C1=O	CCCCCCCCCCCC1=C(O)C(=O)c2cc(O)ccc2C1=O
CCCCCCCCCCCC1=C(C)C(=O)c2ccccc2C1=O	CCC1=C(O)C(=O)c2ccccc2C1=O	Cc1cccc(SC2=CC(=O)c3ccccc3C2=O)c1
CCCCCCCCCCC1=C(C)C(=O)c2cc(Cl)c(Cl)cc2C1=O	CCCCCC1=C(O)C(=O)c2ccccc2C1=O	CC1=C(Sc2ccc3ccccc3c2)C(=O)c4ccccc4C1=O
Cc1cccc2C(=O)C(=C(C)C(=O)c12)C	CCCCCCC1=C(O)C(=O)c2ccccc2C1=O	
Cc1cc2C(=O)C(=C(C)C(=O)c2cc1C)C	CCCCCCCC1=C(O)C(=O)c2ccccc2C1=O	

Table 3.3 General organic molecule SMILES in the dataset

c1ccc2ccccc2c1	COc1cccc(N=Cc2ccccc2)c1	CC=Cc1ccc(C)cc1	CC1=CCCCC1	COc1cccc(N)c1
	C(=NN(c1ccccc1)c1ccccc1)c1ccccc1	O=C(OC(=O)c1ccccc1)c1ccccc1	C=Cc1ccccc1	c1ccc(N2CCOCC2)cc1
CC=Cc1ccc2c(c1)OCO2	Cc1ccc(N=Cc2ccccc2)cc1	O=C1C=CC(=O)O1	C=C(C)C=C(C)C	C1CCNC1
c1ccc2occc2c1	COc1cccc(N=Cc2ccccc2)c1	O=[N+]([O-])c1ccccc1S(=O)(=O)Cl	CC=C(C)C	c1ccc(Nc2ccccc2)cc1
c1ccoc1	Fc1ccc(N=Cc2ccccc2)cc1	O=S(=O)(Cl)c1cc(C(F)(F)F)cc(C(F)(F)F)c1	Cc1cc(C)cc(C)c1	C1CCNCC1
c1ccc2ncccc2c1	C(=Nc1ccccc1)c1ccccc1	O=C(OC(=O)C(F)(F)F)C(F)(F)F	CC(C)=CCO	Nc1ccccc1
c1cn[nH]c1	CC=Cc1ccccc1OC	O=C1CCC(=O)N1Cl	C1=CC2CCC1C2	CC(C)NC(C)C
Clc1ccc(I)cc1	ON=Cc1ccccc1	CC=Cc1cccc(OC)c1	C1=CCCC1	C1CCC2=NCCCN2CC1
Brc1ccc(I)cc1	O=C1CCCCC1	O=C(O)c1ccccc1-c1c2cc(Br)c(=O)c(Br)c-2oc2c(Br)c(O)c(Br)cc12	C1=CCCCC1	COc1ccc(S)cc1
CC=Cc1ccc(OC)cc1	CC(=O)c1ccc(C)cc1	c1ccc2cc3ccccc3cc2c1	C=C(C)CCC	Sc1ccc(S)cc1
O=C(CBr)c1ccccc1	CC(=O)c1ccc(F)cc1	c1ccc(N=Nc2ccccc2)cc1	Oc1ccc(O)cc1	Cc1ccccc1
O=[N+]([O-])c1ccc(F)cc1	CC(=O)c1ccccc1	O=C(c1ccccc1)c1ccccc1	COc1ccc(O)cc1	COc1ccc(SSc2ccc(OC)cc2)cc1
CCOC(=O)CBr	CC(=O)c1ccc(Cl)cc1	O=[N+]([O-])c1ccccc1	Cc1cc(C)c(O)c(C)c1	Sc1ccc2ccccc2c1
CCOC(=O)C(Br)C(=O)OCC	CC(=O)c1ccc(-c2ccccc2)cc1	C1=C(c2ccccc2)CCCC1	Oc1ccc2ccccc2c1	Cc1cccc(C)c1S
Brc1ccc(I)cc1	CC(=O)c1ccc(Br)cc1	N#CC(C#N)=c1ccc(=C(C#N)C#N)cc1	COc1ccccc1O	CSc1ccccc1
Clc1ccc(I)cc1	CC(=O)c1ccc(I)cc1	N#CC(C#N)=c1ccc(=C(C#N)C#N)cc1	Cc1ccccc(C)c1	Cc1ccc(SSc2ccc(C)cc2)cc1
Ic1ccccc1	CC(=O)c1ccc(C(F)(F)F)cc1	C1=CSC(=C2SC=CS2)S1	Cc1ccc(O)cc1C	Sc1ccccc1
Brc1ccccc1	C=C(c1ccccc1)c1ccccc1	C1=CSC(=C2SC=CS2)S1	Oc1ccccc1	c1ccc(SSc2ccccc2)cc1
Clc1ccccc1	CC(=O)c1ccc([N+](=O)[O-])cc1	C#Cc1ccc(OC)cc1	Oc1ccc(Br)cc1	O=[N+]([O-])c1ccc([S+]=[S+]c2ccc(N([O-])[O-])cc2)cc1
O=CC1CCCCC1	CC(=O)C1CCCCC1	CC=Cc1ccccc1C=CC	N#Cc1ccc(O)cc1	S=c1[nH]c2ccccc2[nH]1
CC=Cc1ccc(C=CC)cc1	CN1CCCN(C)C1=O	CC=Cc1ccc(C(C)(C)C)cc1	COc1ccc(OC)cc1	COc1ccc2cc[nH]c2c1
CC(C)CC=O	CN(C)C=O	CC=Cc1ccccc(C=CC)c1	COc1ccc2ccccc2c1	C=Cc1ccc(C)c(C)c1
Cc1ccccc1C=O	O=C1CCCCN1	C1=Cc2ccccc2C1	COc1ccccc1OC	Nc1nccs1
O=Cc1ccccc1	N#Cc1ccccc1	CC=Cc1ccccc1C	CC(C)Oc1cccc(OC(C)C)c1	COc1ccnc2[nH]ccc12
O=Cc1ccc(Cl)cc1	COC(=O)c1ccccc1	CC=Cc1ccc(F)cc1	COc1ccnc(OC)c1	S=c1[nH]c2ccccc2s1
O=Cc1ccc2ccccc2c1	c1ccccc1	CC=Cc1cccc(C)c1	C1=COCCC1	COc1ccc2[nH]ccc2c1
O=Cc1ccc(-c2ccccc2)cc1	O=C(O)c1ccccc1	CC=Cc1ccc(Cl)cc1	Cc1ccccc1C	Cn1cccc1
O=Cc1ccc(C(F)(F)F)cc1	O=C1CCCO1	CC=Cc1ccccc1	COc1ccccc1	S=c1[nH]c2ccccc2o1
N#Cc1ccc(C=O)cc1	N#Cc1cccc(C#N)c1	CC=Cc1ccc(Br)cc1	c1ccc(Oc2ccccc2)cc1	c1c[nH]cn1
O=Cc1ccc([N+](=O)[O-])cc1	COC(=O)c1ccc(C#N)cc1	C(#Cc1ccccc1)c1ccccc1	C1=C(N2CCCC2)CCCC1	c1ccc2[nH]ccc2c1
O=Cc1ccc(O)cc1	N#Cc1ccccc1C#N	CC1=CCCC1	Nc1ccccc1N	c1ccc2sccc2c1
CC=Cc1ccsc1	N#Cc1ccc(C#N)cc1	CC=Cc1ccccc(Cl)c1	C1=C(N2CCOCC2)CCCC1	OCc1ccco1
COc1ccc(C=O)cc1	Cc1cc(C)c(S(=O)(=O)Cl)c(C)c1	C=C(C)c1ccccc1	CN(C)c1ccccc1	
Cc1ccc(N=Cc2ccccc2)cc1	Cc1ccc(S(=O)(=O)Cl)cc1	CC=Cc1cccc(Br)c1	Cc1ccc(N)cc1	
C(=Nc1ccccc1)c1ccccc1	O=C1OC(=O)c2ccccc21	CC=Cc1ccc(C(=O)OC)cc1	CC(C)c1ccccc(C(C)C)c1N	
Fc1ccc(N=Cc2ccccc2)cc1	O=C(Cl)c1ccccc1	CC=Cc1ccccc1Cl	C#Cc1ccccc1	

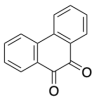
Index	Molecule	Structure	Type
1	1,4-benzoquinone		Benzoquinone
2	Tetramethyl-1,4-benzoquinone		Benzoquinone
3	Tetrachloro-p-benzoquinone		Benzoquinone
4	2,3-Dichloro-5,6-dicyano-1,4-benzoquinone		Benzoquinone
5	juglone		Naphthoquinone
6	menadione		Naphthoquinone
7	1,4-naphthoquinone		Naphthoquinone
8	chimaphilin		Naphthoquinone
9	Dichloronaphthoquinone		Naphthoquinone
10	Dibromonaphthoquinone		Naphthoquinone
11	2,3-dichloro-5,8-dimethoxy-1,4-naphthoquinone		Naphthoquinone
12	anthraquinone		Anthraquinone
13	9,10-Phenanthrenequinone		Anthraquinone
14	1,10-Phenanthroline-5,6-dione		Anthraquinone

Figure 3.3 Li experimental dataset molecular structures

3.2.2 Feature evaluation

The quinone datasets follow the same fundamental rules for the reduction of the basic quinone structure:

- One carbonyl oxygen is reduced to a singly bonded oxygen anion
- The corresponding carbonyl carbon houses the electron previously involved in the carbon-oxygen π bond

The reduction reactions for the general organic compounds followed specific reduction rules:

- A given atom cannot exceed 8 valence electrons
- Triple bonds are broken before double bonds,
- π-bonds are reduced to radicals before sigma bonds,
- The (-) charge from reduction are placed in a position closest to the bond that is broken,
- Electrons from reductions will prefer be housed on atoms of comparatively greater electronegativity within the same molecule
- Bonds are reduced in an order that preserves the maximum number of aromatic π-bonds and retains the stable aromatic structure of the compound.

We trained a GPR model on each molecular feature individually as well in combination. So, in all, four individual features and four combinations were used to train the models, for each of the three datasets. Five-fold random splits were used to train the model and report average results across the splits.

Evaluation of individual features

For this evaluation, we trained four GPR models using Morgan fingerprints, molecular descriptors, DFT energy values, or ADME properties as features. We report the performance on the test set as R^2 (Table 3.4A). Molecular descriptors show the best performance across all datasets including the Li test set (R^2 = 0.600 ± 0.217 - 0.828 ± 0.169 on training dataset, R^2 =

0.842 for Li test set), demonstrating the broad applicability of molecular descriptors for redox potential prediction. Among the many properties captured in molecular descriptors relevant to redox potential, the Partial Equalization of Orbital Electronegativities [113] method used to calculate the Van der Waals surface area of atoms with partial charge captures information about the molecular weight and the number of EDGs and EWGs. Molecular descriptors also capture several molecular properties such as the number and types of functional groups, aromatic rings, etc. Such structural properties play a role in electron donor and acceptance capabilities of the molecules and hence their redox potential. DFT energies are the smallest descriptor with just two energy values. HOMO and LUMO energies directly correlate to a molecule's ability to donate or accept electrons. DFT descriptors perform well on two out of the three datasets (benzoquinones and general organic compounds), but are several orders of magnitude more expensive to calculate as compared to other descriptors. The calculation time for DFT energies is in the order of a few hours as compared to a few seconds for other descriptors. ADME properties also perform well on two out of the three datasets (benzoquinones and naphthoquinones). ADME properties show a low R^2 (0.809) on the benzoquinone dataset as compared to molecular descriptors and DFT energy based descriptors. Morgan fingerprints are purely structural features with no functional properties, and therefore seem to be the least effective in predicting redox potential.

Evaluation of feature combinations

Next, we combined individual features to further improve the performance of our model. Four GPR models were trained on all combinations that utilize molecular descriptors, ADME properties and DFT energy values (Table 3.4B). Since Morgan fingerprints are the least effective individually across all datasets, they were not considered for the combination experiments. We observed that a combination of the remaining three features (molecular descriptors, ADME properties and DFT energies) shows high performance consistently across all training datasets. Each of these features captures some unique insight about the

Table 3.4 Average R^2 for the 10% held out test set for different datasets for different molecular features across 5 random splits. The standard deviation across the 5 splits is also shown for (A) individual features, (B) feature combinations.

Feature	R^2		
	Benzoquinones	Naphthoquinones	General Organic
	(A) Individual features		
Morgan fingerprints	0.113 (± 0.186)	0.052 (± 0.698)	0.438 (± 0.082)
Molecular descriptors	0.828 (± 0.169)	0.655 (± 0.257)	0.600 (± 0.217)
DFT	**0.857** (± 0.234)	0.408 (± 0.428)	**0.721** (± 0.111)
ADME	0.809 (± 0.225)	**0.869** (± 0.073)	0.626 (± 0.217)
	(B) Combination of features		
Descriptors + ADME	0.829 (± 0.177)	0.842 (± 0.064)	0.618 (± 0.218)
DFT + ADME	**0.860** (± 0.198)	**0.867** (± 0.061)	**0.643** (± 0.192)
Descriptors + DFT	0.836 (± 0.171)	0.675 (± 0.262)	0.597 (± 0.224)
Descriptors + ADME + DFT	0.836 (± 0.179)	0.844 (± 0.061)	0.623 (± 0.210)

molecule making their combination the most useful for redox prediction. The next best result is using a combination of molecular descriptors and ADME properties. We also observe that DFT energy features add little to the overall performance considering the computational effort required for calculating these features. The average runtime for running the DFT calculations for each dataset was 3.5 days.

3.2.3 Experimental evaluation

We compared the performance of our model against the experimentally measured redox potentials of 14 molecules - four benzoquinones, seven naphthoquinones, and three anthraquinones (Figure 3.3). The predicted values were adjusted by 500 mV because of the difference in electrodes in the setup for the training set [22] and the Li test set [10]. We investigated the performance of the redox prediction models trained on benzoquinones and naphthoquinones. We did not consider the model trained on the OROP dataset as the Li data was measured under different experimental conditions.

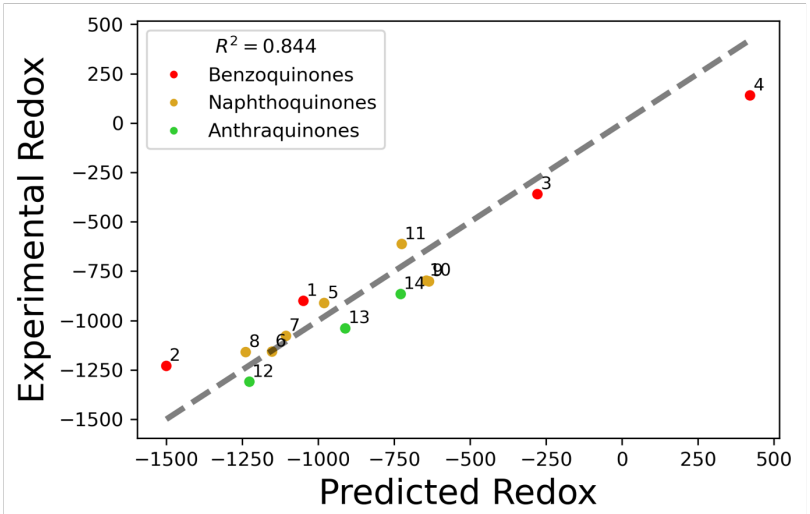
The model trained on the benzoquinones was evaluated on the benzoquinones, naphtho-

quinones, and anthraquinones subsets from the Li dataset. The R^2 values for each subset are 0.835, 0.709, and 0.576, respectively. As expected, the performance on the subset of benzoquinones is the highest followed by naphthoquinones and then anthraquinones. Examining the structure of the compounds (Figure 3.4A), we can also see that the size of the conjugated compounds increases going from benzoquinone, to naphthoquinones to anthraquinones making the redox prediction harder since there are more places for the redox reaction to occur. The resulting R^2 for the entire Li dataset is 0.844 (Figure 3.4B). The R^2 on the entire dataset is higher than on individual subsets since on smaller datasets, one outlier can throw off the R^2 for that subset, but the effect of a single outlier gets mitigated over the larger, complete dataset. Similarly, the model trained on the naphthoquinones was evaluated on the three subsets. The R^2 values for each subset are, -0.039, -0.279, and 0.348, respectively. The R^2 for the complete Li dataset is 0.201 (Figure 3.4C). Unlike the model trained on the benzoquinones, this model does not generalize well on the benzoquinone and anthraquinones. The performance is also poor on naphthoquinones due to one particular naphthoquinone, 2,3-dichloro-5,8-dimethoxy-1,4-naphthoquinone (11 - DCDMNQ). The model performs poorly as it predicts a redox value of -1114.54 mV as against an experimental value of -612.5 mV. For the same molecule, the benzoquinone model performs well predicting a redox value of -725.26 mV. Removing this molecule from the naphthoquinone set changes the R^2 from -0.279 to 0.422 and improves the overall R^2 to 0.308 when using the model trained on the naphthoquinones. The naphthoquinone library is dominated by alkyl substitutions, which would lead to bad predictions of the chlorine and methoxy substitutions in 11. Indeed, the model trained on the benzoquinones better generalizes to other quinones compared to the model trained on naphthoquinones. We also observe that the combination of the molecular descriptors, ADME and DFT features shows the best performance on the complete Li test set (Table 3.5).

(A)

Index	Molecule	Structure	Type
1	1,4-benzoquinone		Benzoquinone
2	Tetramethyl-1,4-benzoquinone		Benzoquinone
3	Tetrachloro-p-benzoquinone		Benzoquinone
4	2,3-Dichloro-5,6-dicyano-1,4-benzoquinone		Benzoquinone
5	juglone		Naphthoquinone
6	menadione		Naphthoquinone
7	1,4-naphthoquinone		Naphthoquinone
8	chimaphilin		Naphthoquinone
9	Dichloronaphthoquinone		Naphthoquinone
10	Dibromonaphthoquinone		Naphthoquinone
11	2,3-dichloro-5,8-dimethoxy-1,4-naphthoquinone		Naphthoquinone
12	anthraquinone		Anthraquinone
13	9,10-Phenanthrenequinone		Anthraquinone
14	1,10-Phenanthroline-5,6-dione		Anthraquinone

(B)



(C)

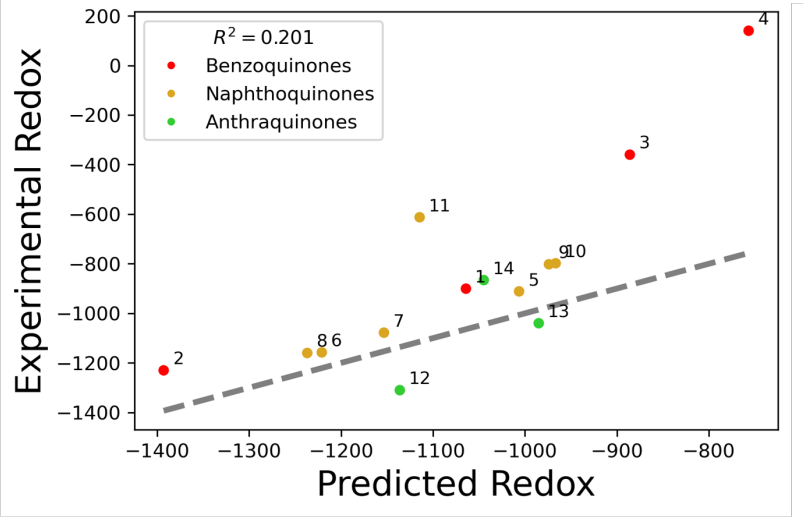


Figure 3.4 Experimental results of the Li dataset evaluated on different models. (A) List of the molecules in the Li test set, (B) All molecules evaluated on model trained on benzoquinones, (C) All molecules evaluated on model trained on naphthoquinones

Table 3.5 Performance of both models with various feature combinations on the Li test set. Morgan fingerprints were not tested because of their poor performance during training

Feature	R^2	
	Benzoquinones	Naphthoquinones
	(A) Individual features	
Molecular descriptors	0.842	-0.049
DFT	-0.880	-0.009
ADME	0.767	-0.304
	(B) Combination of features	
Descriptors + ADME	0.823	0.273
DFT + ADME	0.474	-0.273
Descriptors + DFT	0.840	-0.039
Descriptors + ADME + DFT	**0.844**	**0.201**

3.3 Conclusion

The goal of this study was to test the effectiveness of various types of molecular features in predicting redox potential across three datasets (benzoquinones, naphthoquinones and general organic) while employing the GPR model to predict redox potential. We evaluated four different features individually (Morgan fingerprints, molecular descriptors, DFT energies and ADME properties) and in combination. We showed that the ADME descriptors perform the best individually across all the datasets. We demonstrated that the combination of molecular descriptors, DFT energies and ADME properties provides the best performance across all datasets including the experimental dataset. We achieved relatively high accuracies (R^2 = 0.623 to 0.844) for all three datasets with the best feature combination. We also measured the performance of our model against an experimental dataset of quinones, and we were able to show R^2 = 0.844 for this experimental dataset. With this work, we have proposed an efficient GPR model with a novel combination of molecular features that can be used to rapidly predict redox potential for any given set of organic molecules. Our model can be used to predict redox potentials for large compound libraries. We envision several applications from these redox predictions. Property based molecule synthesis methods

can be conditioned on redox potential as a property. Insights into drug metabolism can be obtained by studying the predicted redox potential of target drug molecules. In terms of further improvements, our model works well on small datasets, and the model can perform even better if trained on a larger set of experimental redox potential values. As larger datasets become available, deep learning methods can also be explored to predict redox potential.

Chapter 4 - CSI: Contrastive Data Stratification for Interaction Prediction and its Application to Compound-Protein Interaction Prediction

Contrastive **S**tratification for **I**nteraction Prediction (**CSI**) is used to stratify (partition) a dataset in a manner that can be exploited via Contrastive Multiview Coding (CMC) [36] to learn embeddings which maximize the mutual information across congruent data views. We showcase the effectiveness of CSI by applying it to the compound-protein sequence interaction prediction problem. CSI partitions the compound-enzyme data by using compound or enzyme as the key. We test our model on 3 datasets against a baseline which does not use contrastive learning.

4.1 Methods

4.1.1 Method overview

The core idea in CSI is intuitive. Each data point is assigned a "key" and multiple views. When learning molecular representations, each "key" is the molecule itself, and the corresponding views are the molecule and a set (or subsets) of interacting sequences. Similarly, when learning sequence representations, the "key" is the sequence itself, and the corresponding views are the sequence and the set (or subsets) of interacting molecules. When stratifying by interaction feature, the "key" is the interaction feature (e.g., all reactions performing a specific biotransformation such as the addition of carboxyl group), and three

views of each reaction (or reaction group, if the key places multiple reactions within a strata) are readily available: reactant-product pairs associated with the reaction (View 1), compound-sequence pairs (View 2), and sequences catalyzing the reaction (View 3), where the compounds are either reaction substrates or products. Other interaction features can also be selected as keys (e.g., reactions sharing homologous sequences). Views under the same key form congruent views of the data, while views across different keys become non-congruent views. Once congruent and non-congruent data views are established, it is possible to apply any contrastive learning technique to learn the joint representation. In our case, we use Contrastive Multiview Coding (CMC) [36], which simultaneously maximizes the mutual information present among the congruent views of the data while discarding features not shared among the views. Our work demonstrates the importance of view selection when applying contrastive learning [114].

4.1.2 Stratification on interaction data - congruent views of compounds and of protein sequences

An interaction dataset consists of compound-protein pairs known to interact. A compound may interact with multiple proteins, and a protein may have interactions with multiple compounds (Figure 4.1). For data stratified using compounds as the key, the set of all protein sequences which interact with the given compound presents a view congruent with the compound. Assuming a lock-and-key based binding model [115], the rationale for these views being congruent is that the interacting proteins have common features which enable binding with the same compound. Subsets, or even pairs, of the protein sequences therefore offer a view which is congruent with the compound. To simplify our formulation and implementation, we use two sequences as a congruent view of a compound. Increasing that number would result in encoders with a higher number of trainable weights. Assuming that I is the set of known interactions on compounds C and a set of sequences S, the set of

congruent views, V_C, for all compounds in C is:

$$V_C = \{[c, (s_i, s_j)],\ s_i, s_j \in S, c \in C, \forall (c, s_i), (c, s_j) \in I\} \tag{4.1}$$

where the square brackets denote views.

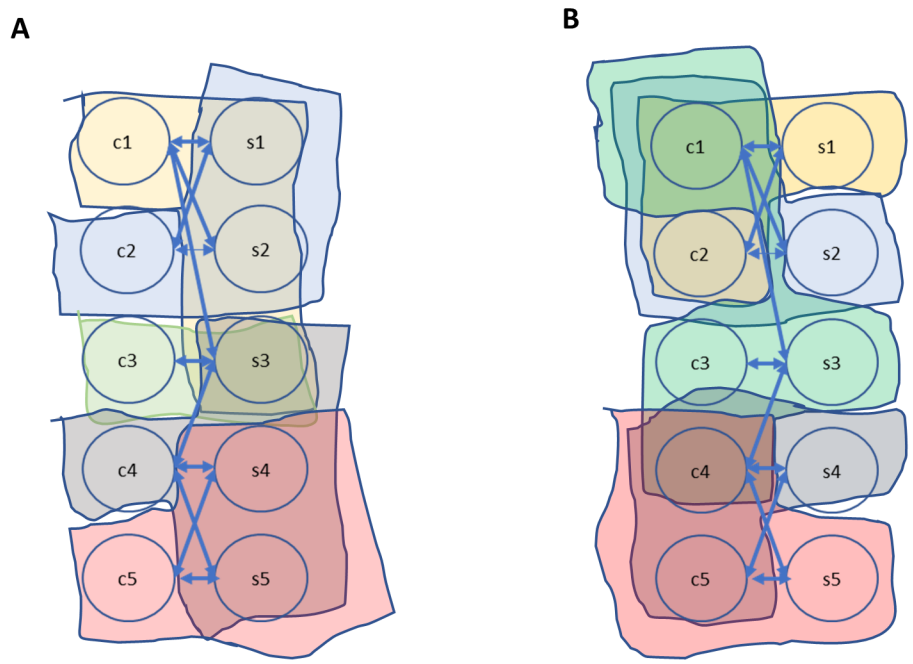


Figure 4.1 Many-to-many interactions between compounds and protein sequences allow data stratification by: (A) compound, and by (B) sequence.

4.1.3 Stratification on reaction data - congruent views on interaction features

Compound-protein interactions within enzymatic datasets are associated with biochemical reactions. The auxiliary data available on the reactions can be used as keys for stratifying by interaction features (and not by compounds and sequences as presented in the prior section). Each reaction represents a set of reactants which undergo a biochemical transformation into a a set of products. Homologous enzyme sequences (e.g., enzymes from different organisms catalyze the same reaction) and multiple enzymes performing similar function can catalyze the same reaction. A biochemical reaction, b, is assumed to be bidirectional and can be

represented as:

$$R \overset{\text{E}}{\leftrightarrow} P \tag{4.2}$$

where R is the set of reactants, P is the set of products, and E the set of enzyme sequences which catalyze the reaction. A reaction can therefore be defined as, $b = \{R, E, P\}$, where the subscripts on R, E, and P are omitted for clarity. Each reaction therefore lends itself to three congruent views: a list of corresponding reactant-product pairs, a list of compound-sequence interactions, where a compound maybe a reactant or a product, and a list of catalyzing sequences. The set of congruent views, V_B, for the set of biochemical reactions, B, is given by:

$$\begin{aligned} V_B = \{[(r_i, p_j), (c_i, s_k), (s_k, s_l)], \forall r_i \in R, p_j \in P, c_i \in R \cup P, s_k, s_l \in E, \\ b = \{R, E, P\}, \forall b \in B\} \end{aligned} \tag{4.3}$$

4.1.4 CSI on interacting objects

CMC [36] arrives at data representations by learning embeddings for each view, and a function, h_θ, which discriminates a congruent pair among a set of non-congruent views based on the learned embeddings. Once the embeddings are learned via encoders, their parameters are frozen and can be used for the downstream task. We adopt a similar methodology for CSI.

CSI is trained in two phases (Figure 4.2). In the first phase, Phase 1A and 1B, we learn embeddings on compound views and, independently, on sequence views. In Phase 1A, for compound as key, each of the congruent views of the compounds, V_c, consist of one compound and two sequences. We therefore train encoders to generate embeddings for these two views, ensuring that they produce same-dimension embeddings. As compounds can be represented as graphs, we utilize a Graph Neural Network (GCN) to encode the compounds:

$$z_{v1,comp} = GCN(c) \tag{4.4}$$

For the protein sequence, we use a 1-dimensional Convolutional Neural Networks (CNN) on the encoded FASTA [116] sequence, F, normalized to a fixed length (e.g., 1000). As we need to learn the representation of two sequences at-a-time to represent the compound, we utilize a Siamese CNN network which uses the same weights. The twin CNNs are trained in tandem on two encoded input sequences and compute the final embedding for the view, z_{v2}. That is,

$$z_{v2,comp} = CNN(F_i) \oplus CNN(F_j) \tag{4.5}$$

where $\oplus$ is the concatenation operation.

In Phase 1B, for sequence as key, congruent views of a sequence, V_s, comprise one sequence, s, and two compounds, c_i and c_j. To learn the embeddings for these views, we utilize a CNN for the encoded sequence, and a Siamese GCN network for the two compounds. That is,

$$z_{v1,seq} = GCN(c_i) \oplus GCN(c_j) \tag{4.6}$$

$$z_{v2,seq} = CNN(s) \tag{4.7}$$

Independent GCNs and CNNs are trained in Phase 1A and Phase 1B.

The discriminator function, h_θ, between embeddings for pairs of n-th and m-th objects from two views $v1$ and $v2$ is defined as in prior work [36] as the cosine similarity between their embeddings modulated by a temperature parameter τ:

$$h_\theta(z^n_{v1}, z^m_{v2}) = exp(\frac{z^n_{v1}.z^m_{v2}}{\|z^n_{v1}\|.\|z^m_{v2}\|}.\frac{1}{\tau}) \tag{4.8}$$

where τ is a hyper-parameter which controls the importance of non-congruent views in pushing the embeddings apart in the latent space. We define the contrastive loss over a batch of size k as:

$$L^{V_1,V_2}_{contrastive} = \frac{1}{k}\sum_{n=1}^{k}\left[-\mathbb{E}[log\frac{h_\theta(z^n_{v1}, z^n_{v2})}{\sum_{m=1}^{k} h_\theta(z^n_{v1}, z^m_{v2})}]\right] \tag{4.9}$$

Defining the contrastive loss in the context of a batch facilitated the CSI implementation and avoided complex strategies to select the non-congruent views [36]. In essence, we select non-congruent views within a batch, instead of considering all possible non-congruent views within the entire dataset. As the contrastive loss $L^{V_1,V_2}_{contrastive}$ treats V_1 as the anchor view and iterates over V_2, it is not symmetrical. We can similarly anchor V_2 and iterate over V_1 to arrive at $L^{V_2,V_1}_{contrastive}$. The total contrastive loss [36], giving equal weight to both views, is then,

$$L(V_1, V_2) = L^{V_1,V_2}_{contrastive} + L^{V_2,V_1}_{contrastive} \tag{4.10}$$

Once the encoders are trained to minimize the loss, their parameters are frozen during Phase 2. The interaction predictor is an MLP neural network which utilizes the learned embeddings for the compound views, and for the sequence views. The interaction predictor is trained on known positive interactions and on negative interactions, which consists of randomly selected compound-sequence pairs. For the contrastive loss (Equation 4.9), views from different keys within a batch (e.g., one compound and two sequences for the compound-based stratification) are taken as non-congruent, while for the final predictor training, randomly selected compound-sequence pairs are taken as negative data. The final predictor $\hat{y}$ is given by,

$$\hat{y} = MLP((z_{v1,comp} \oplus z_{v2,comp}) \oplus (z_{v1,seq} \oplus z_{v2,seq})) \tag{4.11}$$

The prediction loss is the cross entropy loss between $\hat{y}$ and the ground truth y weighted by the ratio of negative to positive labeled data.

4.1.5 CSI on interaction features

When data is keyed by interaction features (Figure 4.3), we apply contrastive loss on three data views: a set of compound-compound pairs representing substrates-products, a set of paired compound-sequences and a set of sequences. The framework of CSI can

be easily adapted to maximize the mutual information across the three views, as was suggested for CMC [36], and to perform interaction prediction on the concatenated learned embeddings. In the first phase of CSI, Siamese GCN and CNN networks are used to learn the compound-compound and sequence-sequence embeddings, and a GCN-CNN are used to learn the embeddings for the compound-sequence embeddings. The contrastive loss is calculated pairwise, over all the views, as defined previously (Equation 4.9). In the second phase, encoder parameters are fixed, and the embeddings from all the neural networks are concatenated and used to train an MLP for interaction prediction.

4.2 Experiments and results

4.2.1 Dataset details

Three datasets, Binding DB, BRENDA, and KEGG, were used to evaluate CSI (Table 4.1A). BindingDB, www.bindingdb.org/bind/chemsearch/marvin/SDFdownload.jsp?download_file=/bind/purchase_target_10000.tsv, provides interaction data for purchasable BindingDB compounds. The KEGG dataset, www.kegg.jp/kegg/download/, is processed to extract the interaction and reaction data. The BRENDA dataset, www.brenda-enzymes.org/download.php, was downloaded as a text file. We used our own tool, PerBRENDA (https://github.com/HassounLab/PER_BRENDA) to process the entries and extract the interaction information for enzymes and substrates.

Binding DB has the highest number of compounds and the highest number of compounds per sequence (9.34 ratio), as expected from a drug-target interaction dataset. For the BRENDA dataset, we extracted interactions between enzymes and ligands as positive interactions. The listed inhibitor interactions were included as labeled negative interactions for the interaction predictor training [66]. For the KEGG dataset, the interactions were extracted from reactions available in the KEGG database. The two enzymatic datasets, the BRENDA and KEGG datasets, have overlap as the BRENDA database covers enzymes interacting with both natural and non-natural substrates, while the KEGG database covers

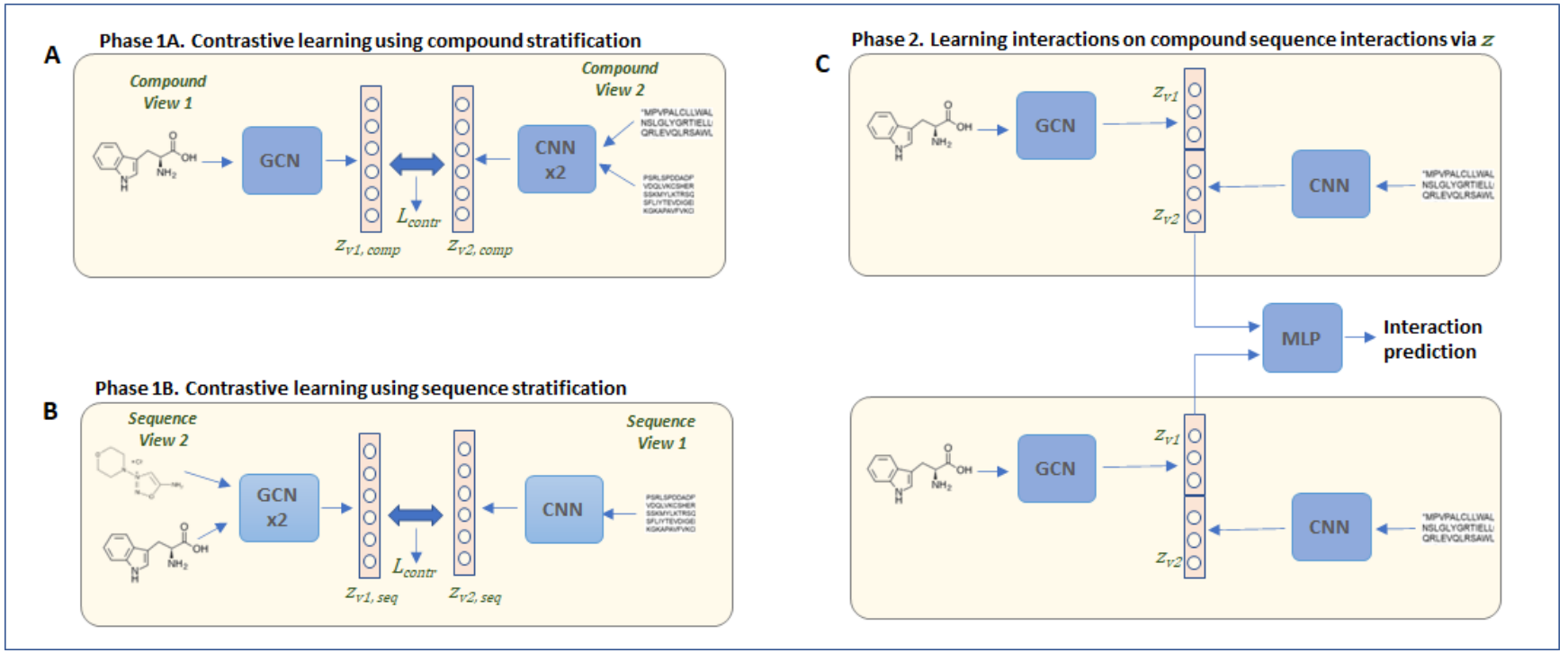


Figure 4.2 CSI model when stratifying each interacting object. (A) Phase 1A for compounds as keys – Compound representation, z_{v1} is generated through a GCN and sequence-sequence representation, z_{v2} is generated using a Siamese CNN. (B) Similarly, in Phase 1B for sequences as keys, compound-compound representation, z_{v1}, is generated through a Siamese GCN, while sequence representation, z_{v2}, is generated through a CNN. (C) In Phase 2, the trained encoders from Phases 1A and 1B are fixed. The representations are concatenated to train an MLP for final prediction.

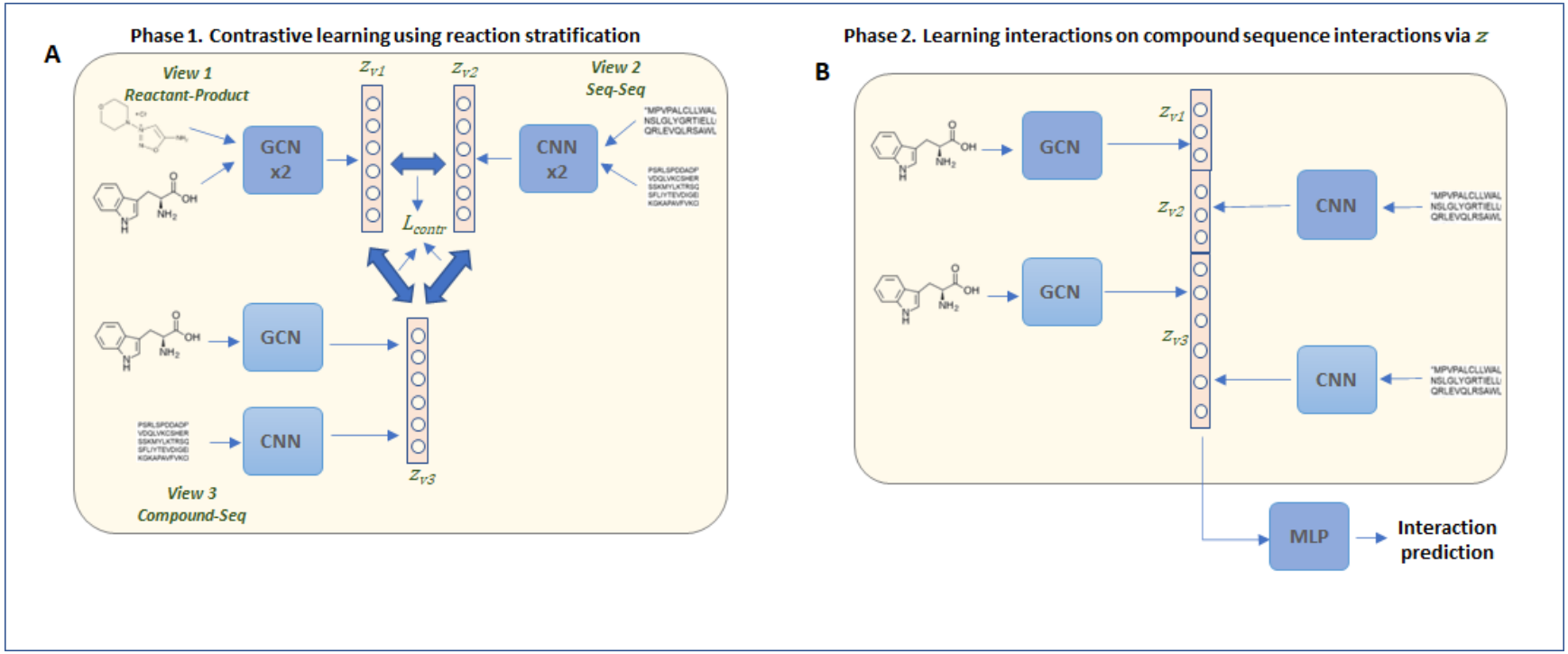


Figure 4.3 CSI model when stratifying by interaction feature. (A) Phase 1. Contrastive loss is applied to the three data views: compound-compound pairs, compound-sequence pairs, and sequence-sequence pairs to generate three embeddings, z_{v_1}, z_{v_2}, and z_{v_3}. (B) Phase 2. Trained encoders from Phase 1 are used to generate representations for compounds and sequences. These representations are concatenated together to train an MLP for the final prediction.

natural interactions found in living organisms. The KEGG database provides detailed information on the underlying biochemical reactions, which enabled stratification on interaction features. The two datasets have 757 compounds which had the same canonical SMILES. Of the 21,367 unique sequences in KEGG dataset, 10,948 are in the BRENDA dataset. The KEGG dataset has approximately $3\times$ more interactions than the BRENDA dataset.

For compound-based stratification (Table 4.1B), we report the size of each strata. Within each strata, the number of views is the square of the number of sequences divided by two as CMC is applied to pairs of sequences within each strata. A large size therefore indicates rich views within the strata. To assess the overlap among the strata, we report the average sharing among the strata and their Jaccard similarities. Th Jaccard similarity metric gives a sense of how varied the views are across keys while also considering the strata size. We similarly summarize these metrics for sequence-based stratification (Table 4.1C). The KEGG dataset has more average shared sequences across compound keys (0.06) compared to the others, while the BindingDB dataset had more average shared compounds across sequence keys (0.14). When considering the Jaccard score, BindingDB has the highest similarities per strata for both compound- and sequence-based stratification.

For the interaction prediction task (Table 4.1D), the training data consists of positive examples comprising protein-compound pairs which are known to interact. The negative examples are randomly selected compound-sequence pairs. The selection strategy of the negatives reflects nature as most compounds and proteins do not interact. For training, we used a negative to positive ratio as 5:1, taking care to appropriately weight the loss during training. We created two kinds of test sets. The **Test** set included both positive and negative examples taken from the same distribution as the training set. We also generated test sets with 5X, 10X and 25X the number of negatives as positives to evaluate the impact of negative-to-positive ratio in test. To test the generalizability of the model, we created an **Unseen Test** set which comprised the 5% least frequent compounds and sequences in each dataset, which were held out from the training data set. We assume a 1:1 negative-to-positive

ratio for the Unseen Test.

Table 4.1 Statistics for the three evaluation datasets. (A) Base statistics. (B) Strata statistics when stratifying by compound. (C) Strata statistics when stratifying by sequence. (D) The number of positive examples for various data splits.

(A) Statistics for the interaction dataset			
	BindingDB	BRENDA	KEGG
Number of interactions	68,347	40,693	127,884
Unique compounds	29,149	8,891	6,087
Unique sequences	3,120	13,330	21,367
Compound-to-sequence ratio	9.34	0.67	0.28
(B) Stratification on compounds (reported in sequences per strata)			
Average strata size	2.18	7.2	21
Standard deviation	6.85	31.7	64.1
Maximum strata size	339	1518	2065
Minimum strata size	1	1	1
Average sharing among strata	0.01	0.01	0.06
Average Jaccard similarity among strata	0.004	0.001	0.001
(C) Stratification on sequences (reported in compounds per strata)			
Average strata size	20.4	3.5	5.9
Standard deviation	47.4	4.1	8.9
Maximum strata size	623	107	331
Minimum strata size	1	1	1
Average sharing among strata	0.14	0.04	0.06
Average Jaccard similarity among strata	0.003	0.001	0.001
(D) Number of positives per data split			
Training	49,746	32,536	100,999
Validation	6,142	4,095	12,626
Test	7,833	4,925	14,261
Unseen Test	1,609	885	1,636

4.2.2 Baseline model

While our proposed data stratification strategy can be applied to any interaction model, we create a baseline model which utilizes Graph Neural Networks (GNNs) to encode the molecules, and Convolutional Neural Networks (CNNs) to encode the sequences (4.4. Compounds represented in SMILES format are converted to a molecular graph using rdkit

[117]. For our baseline, we use node features as the atom type, atomic mass, valence, is atom in ring, formal charge, radical electrons, chirality, degree, number of hydrogens and aromaticity. Bond features are the bond type, whether the bond is part of a ring, conjugacity and one hot encoding of the stereo configuration of the bond. Compound embeddings are learned using a multi-layer Graph Neural Network (GNN) encoder. The network consists of Graph Convolutional Networks (GCNs) [118] which aggregate information at each node. The GCNs are followed by a pooling layer and two fully connected layers. Our baseline is the same as the GraphDTA model [71], which is reported to outperform other state-of-art models such as [119, 120, 121], and thus provides a strong baseline. Since GraphDTA is a regression model which predicts the binding affinity, we modified the final layer of GraphDTA to enable binary prediction as needed for our interaction prediction problem. Compounds represented in SMILES format are converted to to a molecular graph using rdkit [117]. For our baseline, we use node features as the atom type, atomic mass, valence, is atom in ring, formal charge, radical electrons, chirality, degree, number of hydrogens and aromaticity. Bond features are the bond type, whether the bond is part of a ring, conjugacity and one hot encoding of the stereo configuration of the bond. Compound embeddings are learned using a multi-layer Graph Neural Network (GNN) encoder. The network consists of Graph Convolutional Networks (GCNs) [118] which aggregate information at each node. The GCNs are followed by a pooling layer and two fully connected layers. Each amino acid within protein sequences (in FASTA format) is first converted to a numeric code used to generate learnable embeddings. Sequence embeddings are passed to a protein encoder which consists of a 1-d Convolutional Neural Network (CNN) followed by a pooling layer and a fully connected layer. The compound and sequence embeddings are concatenated for the final interaction likelihood prediction. The final predictor is a 3-layer MLP with the first two layers each reducing the embedding dimensionality by half and the final layer making a binary prediction. Importantly, the architecture of the GCN and CNN encoders of the baseline model are used for CSI to ensure a fair comparison between the CSI and baseline

model.

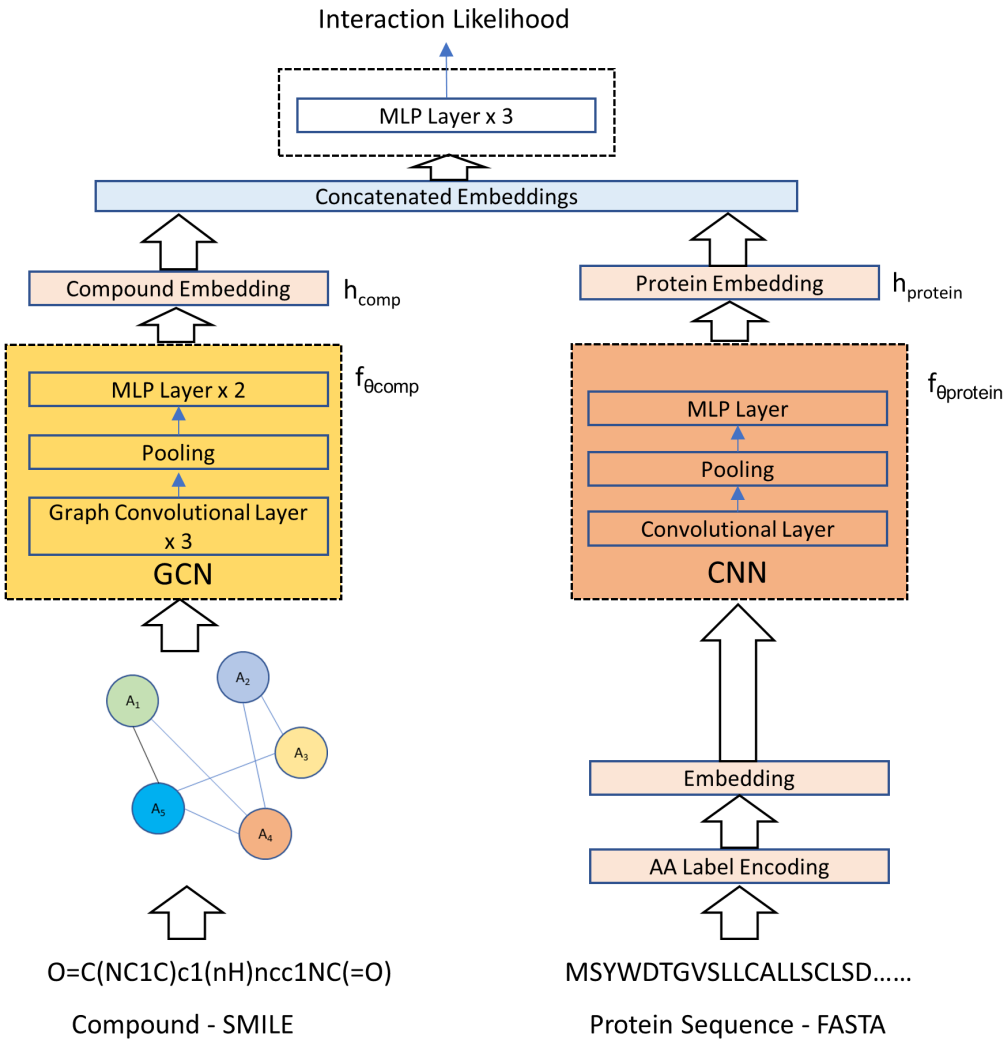


Figure 4.4 Compound-protein interaction prediction model used as the baseline. Interaction likelihood is predicted based on learned molecular and protein interactions.

4.2.3 Experimental setup

To evaluate the CSI model, we measure the model's performance in ranking positive examples ahead of negative examples, as well as the model's ability to rank a molecule or sequence which has the highest probability of interacting with a sequence or molecule respectively. We used Average Precision, Mean Average Precision and R-Precision as the

metrics, the details of which are available in Section 2 of Supplementary material.

The following metrics were used to measure model performance:

- **Average Precision (AP)** is measured across each dataset, reflecting the model's ability to distinguish positive and negative examples.
- **R-precision**, also measured across each dataset, measures the model's ability to accurately predict the R known positive interactions.
- **Mean Average Precision (MAP)** measures the the AP per compound (or per protein sequence) averaged over interactions sorted by compound (or protein sequence), thus indicating the model's ability to predict the likelihood of interaction for a given compound (or protein sequence).
- **MAP@3** reports MAP on the top 3 ranked items i.e the top 3 sequences per compound, or the top 3 compounds per sequence.
- **Precision@1** measures the ability of the model to correctly predict a top ranked interacting item.

The CSI model was trained in two steps. In the contrastive learning step, the encoders generating z_{v1} and z_{v2} were trained using CMC on the congruent and non-congruent data views. The model was trained for 700 epochs. The best temperature τ was found to be 0.07 (we tried a range of 0.05-0.08). Adam [122] was used as the optimizer. In the interaction prediction step, the training set was divided into training, validation and test sets in ratio 8:1:1. In this step, the predictor model was trained for 200 epochs, with early stopping on validation loss. The optimizer used was Adam.

4.2.4 Results on stratification by compounds and sequences

The results for the three datasets are reported for test set with a negative-to-positive ratio of 1:1 (Table 4.2A). The CSI model shows improved performance across all datasets and

across all metrics. CSI significantly outperforms the baseline model which does not use CSI, where average precision (AP) is improved by 18.2% on the BindingDB dataset, 39% on the BRENDA dataset, and 13.7% on the KEGG dataset, when stratifying by compound and by sequence. The improvement in MAP over baseline is maximum on the BindingDB dataset (23.8%) for test data sorted by sequences while it is maximum on KEGG (26.2%) for test data sorted by compounds. For the Unseen Test set, CSI also shows improved performance across all metrics, and across all datasets (Table 4.2B). Specifically, AP improvements are 2.6%, 18.2% and 1.6% for BindingDB, BRENDA and KEGG datasets, respectively. Clearly the quality of these datasets are different, and hence the performance. The richness of the strata within each dataset, measured by the variety (i.e. less sharing) across views, impacts how contrastive learning performs on each dataset.

The CSI model uses embeddings learnt based on compound and sequence stratification. To determine which of the two stratification strategies contributes more to performance gains over the baseline model, we performed ablation studies on BindingDB (non-enzymatic) and KEGG (enzymatic) datasets (Table 4.2C). Independently, compound and sequence stratification each contribute significantly to CSI's performance over the baseline. For BindingDB, with sequence based stratification alone, the AP drops from 0.992 (with both stratification) to 0.972. The drop is minimal (0.992 to 0.991) when using only compound-based stratification. These results indicate that for BindingDB, the compound based stratification contributes maximally to the CSI model's performance. For KEGG, the AP drops from 0.969 when using both stratification strategies to 0.953 when using only compound-based stratification. The drop is lesser when switching to sequence based stratification (0.969 to 0.960), indicating that for KEGG, the sequence-based stratification contributes maximally to the CSI model's performance. The performance of the CSI model also scales well when the ratio of negative to positive examples in increased to mimic what happens in nature.

Table 4.2 Interaction prediction results for a negative data ratio of 1:1 for the baseline (GraphDTA with a binary predictor instead of a regressor) and CSI models for the BindingDB, BRENDA and KEGG datasets. AP and R-Precision are reported for the entire dataset. MAP, mean R-Precision, MAP@3 and R-Precision@1 are reported for data sorted by compounds and by sequences. (A) Test set. (B) Unseen Test. (C) Ablation study to determine the individual contributions of each stratification strategy against using both strategies together.

		Overall		Compound				Sequence			
		AP	R-Precision	MAP	R-Precision	Map@3	Precision@1	MAP	R-Precision	Map@3	Precision@1
(A) Test set											
BindingDB	GraphDTA	0.839	0.783	0.914	0.844	0.913	0.842	0.802	0.709	0.799	0.681
	CSI	0.992	0.971	0.996	0.993	0.996	0.993	0.993	0.993	0.994	0.993
BRENDA	GraphDTA	0.778	0.713	0.804	0.680	0.804	0.680	0.874	0.803	0.874	0.791
	CSI	0.991	0.970	0.995	0.990	0.996	0.993	0.978	0.993	0.978	0.975
KEGG	GraphDTA	0.852	0.770	0.755	0.627	0.739	0.610	0.757	0.857	0.755	0.701
	CSI	0.969	0.902	0.953	0.918	0.956	0.939	0.809	0.968	0.808	0.793
(B) Unseen Test											
BindingDB	GraphDTA	0.975	0.918	0.995	0.991	0.995	0.991	0.920	0.879	0.911	0.879
	CSI	1.000	0.992	1.000	0.999	1.000	0.999	0.997	0.995	0.997	0.995
BRENDA	GraphDTA	0.845	0.703	0.934	0.891	0.927	0.891	0.901	0.857	0.897	0.841
	CSI	0.999	0.985	1.000	1.000	1.000	1.000	0.982	1.000	0.982	0.982
KEGG	GraphDTA	0.873	0.771	0.779	0.682	0.753	0.676	0.730	0.963	0.730	0.718
	CSI	0.886	0.773	0.915	0.878	0.908	0.869	0.718	0.934	0.717	0.697
(C) Ablation study											
BindingDB	CSI Seq Strat	0.972	0.921	0.982	0.973	0.978	0.969	0.971	0.960	0.983	0.982
	CSI Comp Strat	0.991	0.971	0.987	0.989	0.996	0.991	0.983	0.982	0.982	0.982
KEGG	CSI Seq Strat	0.960	0.894	0.942	0.901	0.952	0.931	0.807	0.958	0.808	0.791
	CSI Comp Strat	0.953	0.882	0.923	0.869	0.931	0.900	0.798	0.951	0.802	0.778

4.2.5 Scaling of model performance with increasing negative to positive ratio

We measured model performance scaling with respect to the negative-to-positive ratio (Figure 4.5). Assuming the performance with 1:1 ratio to be 1.0, baseline model AP performance drops by 73% when using 25:1 ratio. Meanwhile, the performance of the CSI model drops only by 18% at the 25:1 ratio. For metrics measured per compound, the MAP metric drops by 75% for the baseline model for the 25:1 ratio whereas the same metric drops by only 14% for the CSI model. For metrics measured by sequence, the drops in MAP are 66% and 13% for the baseline and CSI model respectively. These results indicate that CSI performs better than the baseline model at predicting the negatives correctly. Further, the positives and negatives continue to be well separated even as the ratio of negatives increases.

4.2.6 Results on stratification by reaction features

For the KEGG dataset, three interaction features were used to produce three stratification strategies. The first strategy partitions the data based on enzymes catalyzing the same reaction (e.g., homologs). The second strategy divides the interactions by the underlying biotransformation pattern associated with the substrate-product pairs. KEGG classifies reactions based on this property, and each class is referred to as an RCLASS [123]. Multiple reactions can belong to the same class and result in similar biotransformations. The third strategy divides the interaction data by the Enzyme Commission (EC) number associated with the interaction. EC numbers provide hierarchical classification on enzymes and are represented as four numbers separated by periods (e.g. L-lactate dehydrogenase is assigned EC number 1.1.1.27). Each such EC number is associated with one or more biochemical reactions. The three keys used to partition the KEGG interaction data are therefore: the reaction, RCLASS, and the EC numbers. For each strata (Table 4.3), three different views of the data are possible: substrate-product pairs, compounds-sequence pairs and pairs of sequences. The number of keys per strategy differ, where stratification on reactions provides the most number of keys. The key choices subsequently affect the total number of views and

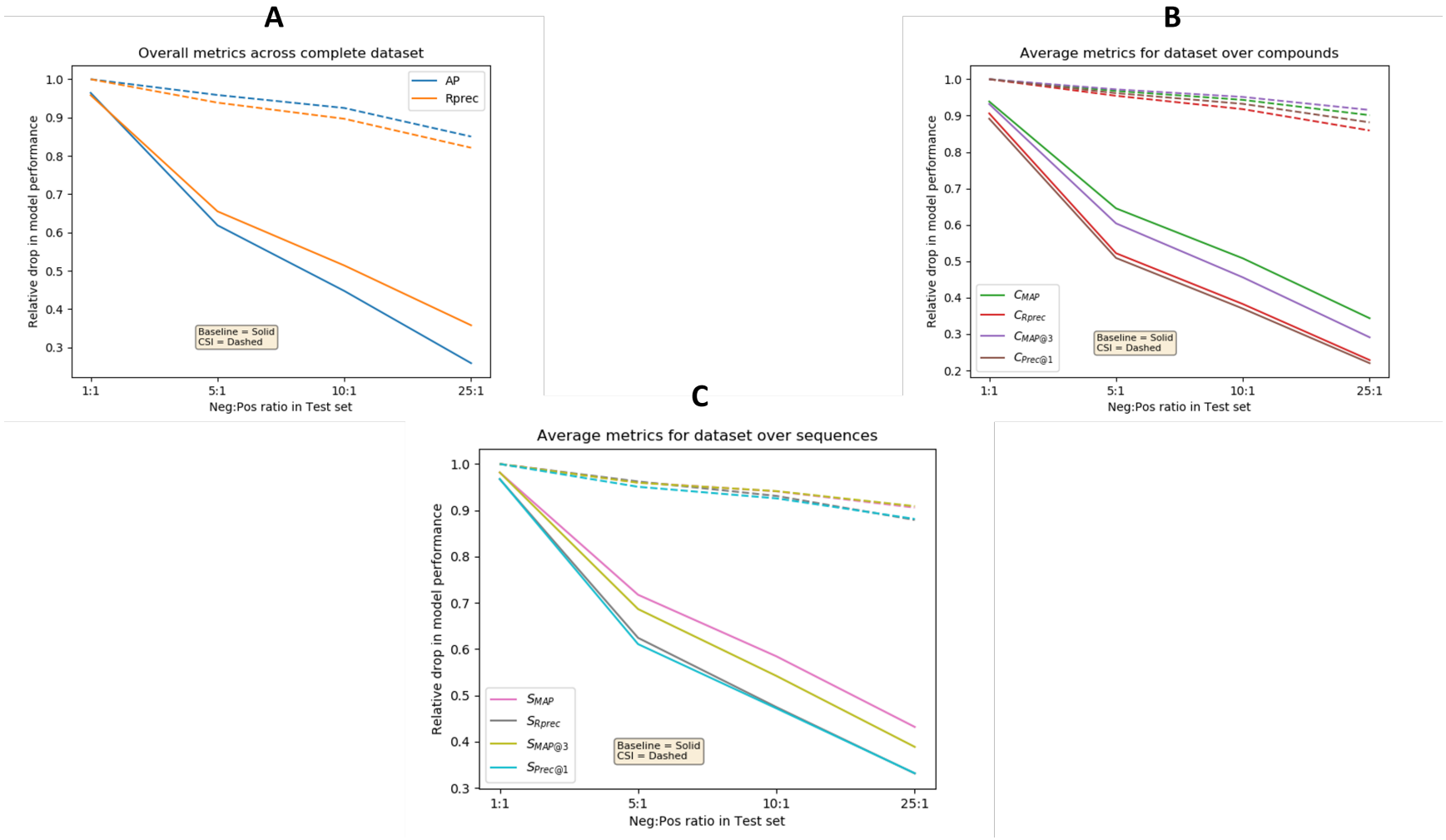


Figure 4.5 Model performance evaluation for various negative-to-positive ratios in the Test set. (A) AP and R-precision trends for various negative-to-positive ratios for Test set. (B) MAP, mean R-Precision, MAP@3, R-Precision@1 trends for Test set interactions sorted by compounds. (C) MAP, mean R-Precision, MAP@3, R-Precision@1 trends for Test set interactions sorted by sequences.

the size of each strata within the views. Regardless of the key, there are more compound-sequence views (V_2) than the other two views, and fewest compound-compound views (V_1). We examine the strata to determine if any one particular reaction consistently contributed to the maximum strata size. The largest compound-sequence partition under the reaction stratification strategy is due to the ammonia-ubiquinol reaction (KEGG reaction R00148), which contributes to nitrogen metabolism. This reaction is catalyzed by ammonia mono-oxygenase (EC 1.14.99.39), which is present in 46 organisms leading to many sequences for the same enzyme. For RCLASS-based stratification, the largest compound-sequence partition is for RC00001 which is part of glutathione metabolism. This reaction class encompasses 14 different reactions, catalyzed by 18 different EC classes - leading to a large number of compound-sequence pairs. For EC-based classification, the largest compound-sequence partition is for glutathione transferase (EC 2.5.1.18), which catalyzes 24 reactions, and present in 423 different organisms.

Table 4.3 Statistics for the KEGG dataset for three different stratification strategies by interaction features. We report the total number of objects in each view with each stratification strategy, the average number of objects in each views over all keys, as well as the distribution of objects in each view.

	Number of views	Mean objects	std-dev objects	Max objects
(A) Stratification on reaction. Number of keys is 6,059				
V_1(comp-comp)	9,091	1.50	0.58	4
V_2(comp-seq)	124,046	20.47	35.32	387
V_3(seq-seq)	52,759	8.71	15.36	129
(B) Stratification on RCLASS. Number of keys is 2,158				
V_1(comp-comp)	5,765	2.67	7.09	134
V_2(comp-seq)	172,382	79.88	441.91	15635
V_3(seq-seq)	50,723	23.50	107.47	2994
(C) Stratification on EC. Number of keys is 3,363				
V_1(comp-comp)	8,493	2.53	3.00	79
V_2(comp-seq)	98,254	29.22	170.72	7500
V_3(seq-seq)	17,563	5.22	8.47	129

The results are reported using AP (Table 4.4) as this metric was well correlated in earlier

analysis with other metrics. AP was reported for the baseline model on three datasets (1:1 negative-to-positive ratio, 5:1 ratio, and the Unseen Test) as well as for the CSI model for the same datasets. All stratification strategies yield improved results over the baseline model and stratification by compound and sequence across all test sets, where stratification by reaction outperforms the baseline by 16.9%, 62.6% and 13% on the 1:1, 5:1, and Unseen Test sets, respectively. Comparing with stratification by compound and sequence, stratification by reaction yields AP improvements over the compound and sequence stratification by 2.1%, 6.3% and 10.8% on the 1:1, 5:1, and Unseen Test sets, respectively.

As expected, the performance of both the baseline and CSI model drops as the negative-to-positive ratio increases, but whereas the baseline model performance drops by 43% between Test 1:1 and Test 5:1, the performance of the CSI model with reaction based stratification drops by 2.7% for the same test sets. Of the three stratification strategies, stratification by reaction provides the most improvement over the baseline model.

The maximum improvement (63% over baseline) was for the test set with a negative-to-positive ratio of 5:1, while the 1:1 test set and Unseen Test set yielded improvements of 16.5% and 12.6%, respectively. In comparison, the earlier compound and sequence based stratification strategy show 54%, 13.7% and 1.5% improvement in performance respectively over the 5:1, 1:1 and Unseen data sets. Clearly, the stratification by interaction feature results in improved performance as compared to compound and sequence based stratification strategies.

To evaluate how each of the views contributes to improvements over the baseline, we perform an ablation study (Table 4.4B). As stratification by reaction resulted in higher performance over RCLASS- and EC-based stratification, the ablation study is applied to the reaction-based stratification model. The model was successively trained on each combination of two views (instead of all 3). Removing V_1 (substrate-product view) contributed the most, when compared to removing the other two views, in reducing model performance, e.g., for the 5:1 positive-to-negative Test set, the AP performance is reduced from 0.963 to 0.751.

The substrate-product view therefore contributes the most to the CSI model performance when stratifying by reaction features. We conjecture that the high similarity between substrate-product pairs contributes to higher mutual information when compared to the other views.

Table 4.4 AP results on stratification for the baselines (no stratification, compound/sequence stratification) and by the three interaction features: reaction, RCLASS, and EC. The three views, V_1, V_2, and V_3, correspond to substrate-product pairs, compounds-sequences and pairs of sequences. The ablation study considers only two of the views at a time.

Model	Test	Test(5:1)	Unseen Test
(A) Summary of prior results			
Baseline (no stratification)	0.852	0.587	0.873
Compound/sequence stratification	0.969	0.906	0.886
(B) Interaction features			
Reaction (V_1, V_2, V_3)	0.989	0.963	0.982
RCLASS (V_1, V_2, V_3)	0.990	0.913	0.954
EC (V_1, V_2, V_3)	0.988	0.943	0.979
(C) Ablation study			
Reaction Strat (V_1, V_2)	0.980	0.874	0.941
Reaction Strat (V_2, V_3)	0.962	0.751	0.904
Reaction Strat (V_1, V_3)	0.983	0.902	0.947

4.3 Conclusion

CSI is a generalizable data stratification technique which exploits relationships among interacting objects to define congruent (and non-congruent) views. Paired with CMC, CSI learns representations which maximize the mutual information among congruent views, leading to enhanced representations for the downstream interaction prediction task. In addition to advancing the state-of-the-art in interaction prediction in the broader field of deep-learning, CSI also advances interaction prediction between protein and molecules, as evidenced by our results. CSI was applied to three compound-protein sequence datasets, involving both enzyme-molecule and protein-target datasets. Our results show significant improvement in AP, in the range of 13.7% to 39% over comparable baselines which do

not utilize stratification. We further demonstrated that, for our datasets, stratification by interaction features results in improved performance over stratification on object relationships. Data stratification as described herein is the new paradigm of "Data-Centric AI" (https://spectrum.ieee.org/andrew-ng-data-centric-ai), where data stratification methods will complement advances in deep learning. A variety of contrastive learning methods, including CSI, have the potential to further advance protein-ligand interaction predictions.

Chapter 5 - JESTR: Joint Embedding Space Technique for Ranking Candidate Molecules for the Annotation of Untargeted Metabolomics Data

A major challenge in metabolomics is annotation: assigning molecular structures to mass spectral fragmentation patterns. Despite recent advances in molecule-to-spectra and in spectra-to- molecular fingerprint (FP) prediction, annotation rates remain low. We introduce in this thesis a novel paradigm (JESTR) for annotation. Unlike prior approaches that *explicitly* construct molecular fingerprints or spectra, JESTR leverages the insight that molecules and their corresponding spectra are views of the same data and effectively embeds their representations in a joint space. Candidate structures are ranked based on cosine similarity between the embeddings of query spectrum and each candidate. We train and test JESTR on three datasets and against two state-of-the-art methods.

5.1 Methods

We address in this paper the problem of assigning chemical structures from a candidate set to spectral data. The novelty of our approach lies in avoiding the explicit generation of molecular fingerprints and spectra (Figure 5.1A), and in considering a molecule and its spectra as views of the same object (Figure 5.1B). This valuable insight allows us to embed molecules and their matching spectra close in the molecule-spectrum joint embedding space.

The JESTR model architecture (Figure 5.2) consists of a molecular encoder and a spectral encoder. They are trained to create embeddings in a molecule-spectrum joint embedding space. To place views of the same object close to each other in the embedding space, we use the CMC contrastive learning loss [36]. To improve performance, we utilize regularization. At inference, when provided a candidate set for the query spectrum, the cosine similarity is computed between each candidate and the query spectrum. The candidates are then ranked based on their cosine similarities.

5.1.1 Encoders

The molecular encoder is implemented using a multi-layer Graph Neural Network (GNN) encoder. Molecular structures are encoded as graphs, where node features are include atom type, atomic mass, valence, if the atom is in a ring , formal charge, radical electrons, chirality, degree, number of hydrogens and aromaticity. Edge features are the bond type, whether the bond is part of a ring, conjugacity and one hot encoding of the stereo configuration of the bond. The encoder consists of Graph Convolutional Networks (GCNs) [118] that aggregate information at each node. The GCNs are followed by a pooling layer and two fully connected layers to generate the final molecular embeddings, z_{mol}, for a given molecule graph c:

$$z_{mol} = MLP_{\times 2}(MAXPOOL(GCN(c))) \tag{5.1}$$

To prepare the spectrum for its encoder, peak m/z values of the spectra are discretized into bins that are 1 Da wide. Peaks with m/z values larger than 1000 Da were dropped. The intensity are normalized to a max value of 999 - a common practice in normalizing spectral data (e.g., for the NIST datasets). For multiple peaks falling within the same bin, peak intensities within each bin are summed to generate the overall intensity value for that bin. A 1000-dimension binned vector therefore encodes the spectrum. A log10/3 transformation is applied to this binned vector to ensure that a few peaks and/or a long tail do not dominate the embedding vector. This 1000-dimension encoded vector was passed through a 3-layer

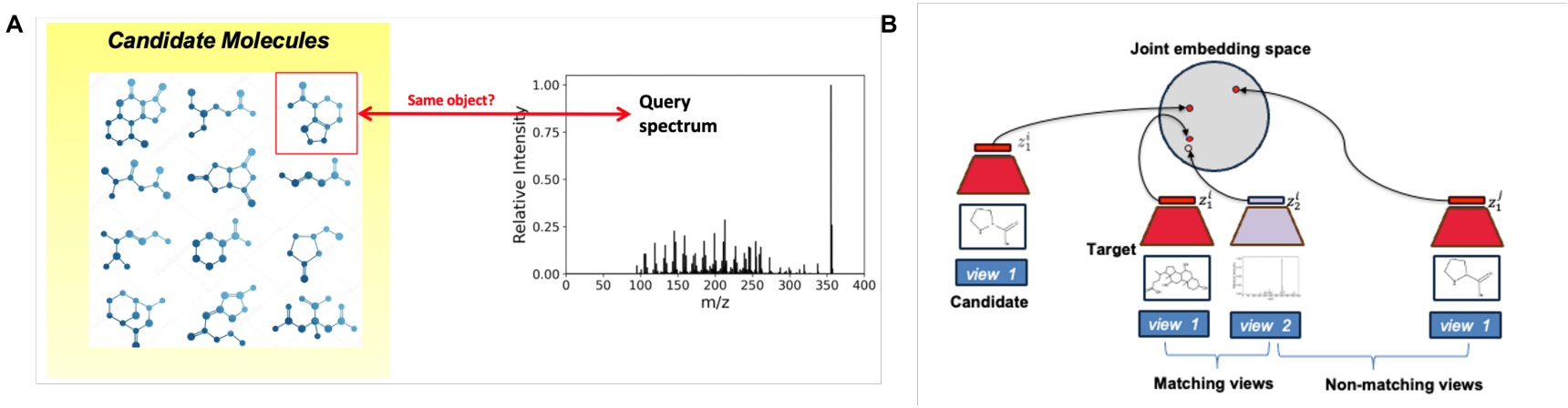


Figure 5.1 Novelty of the JESTR annotation approach. A. JESTR avoids the explicit generation of spectra, molecules, and fingerprints, and ranks the candidate molecules against the query spectrum based on their joint-space embeddings. B. JESTR learns to place representations of matching molecule-spectrum pairs close in the joint embedding space relative to non-matching pairs. Further, JESTR utilizes additional molecules beyond those in the training set to learn to distinguish target molecules in the training dataset from candidate molecules (those with similar molecular formulas).

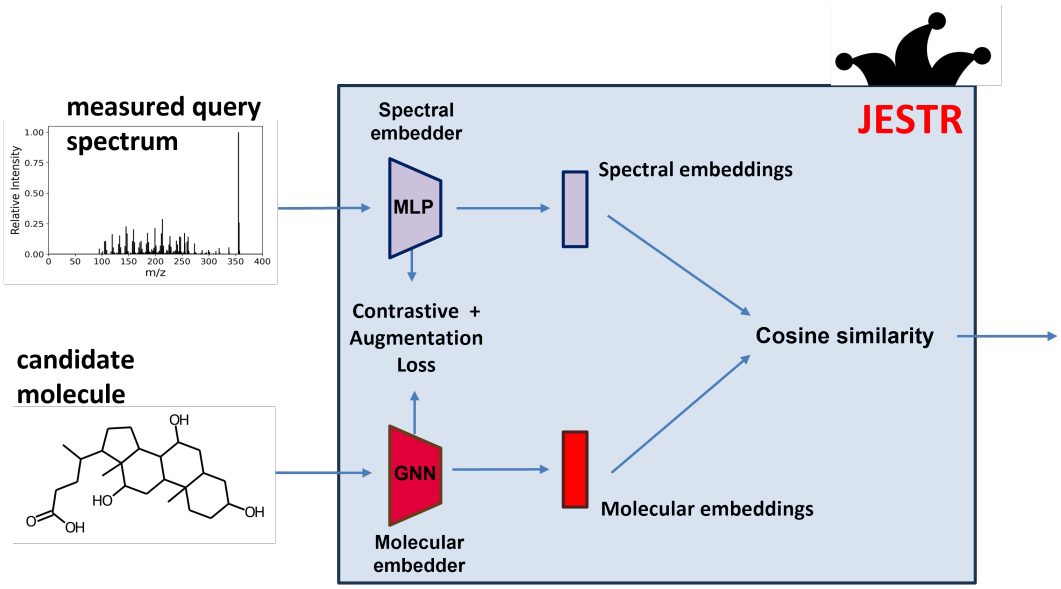


Figure 5.2 Overview of the JESTR model architecture. The model is trained to minimize the contrastive and regularization losses. The embeddings produced by the encoders are used to compute the cosine similarity in the joint embedding space between a molecule and a spectrum.

MLP to obtain the final spectral embedding, z_{spec}:

$$z_s = \frac{1}{3} log_{10}(\{\Sigma\ I_i, \forall\ i,\ n < (mz)_i < (n+1), for\ n\ in\ 0..999\}) \tag{5.2}$$

where I_i is the intensity of the i-th peak and mz_i is the m/z value of the i-th peak.

$$z_{spec} = MLP_{\times 3}(z_s) \tag{5.3}$$

5.1.2 Contrastive learning of spectral and molecular views

We consider two views of each data item: a molecular and a spectral view. Each data item will have one molecular view but may have multiple spectral views as measurements may be collected under different mass spectrometry instrumentation conditions. Matching molecule-spectrum views arise from a molecule and its spectrum, while non-matching views arise between a molecule and any of its non-matching spectra. The objective of contrastive learning on multi-views [37, 38, 36, 39] is to learn embeddings that separate samples from

matching and non-matching distributions, and to ensure that paired views are close in the joint embedding space.

As in CMC [36], we use a discriminator function, h, to measure the closeness of spectral and molecular embeddings using their cosine similarity, modulated by a temperature parameter τ. Thus, given a embedding for spectrum n, and an embedding for molecule m, we can define h as:

$$h(z^n_{spec}, z^m_{mol}) = exp(\frac{z^n_{spec}.z^m_{mol}}{\|z^n_{spec}\|.\|z^m_{mol}\|}.\frac{1}{\tau}) \tag{5.4}$$

The hyper-parameter τ controls the importance of non-matching pairs in pushing the embeddings apart in the joint embedding space. To ensure that the discriminator assigns high values for matching pairs and low values for non-matching pairs, we define a contrastive loss, $L_{contrastive}$, over a batch of size k as:

$$L_{contrastive} = \frac{1}{k}\sum_{n=1}^{k}\left[-\mathbb{E}[log\frac{h_\theta(z^n_{spec}, z^n_{mol})}{\sum_{m=1}^{k} h_\theta(z^n_{spec}, z^m_{mol})}]\right] \tag{5.5}$$

This loss effectively ensures that the cosine similarity between each matching pair, (z^n_{spec}, z^n_{mol}), is highest among all possible pairings, (z^n_{spec}, z^m_{mol}), within the batch.

5.1.3 Regularization

As candidate molecules typically have the same molecular formula as the target molecule, we fetch such candidates from the PubChem database, and utilize regularization to train the model to better distinguish between target molecules and their candidates. Our regularization objective is therefore to push candidates away from the spectra, and hence from the corresponding target, in the joint embedding space. Training using regularization is implemented by introducing an additional loss to minimize the cosine similarity between the embeddings of each spectrum and the candidate molecules of the corresponding target. Candidate sets are sorted by their Tanimoto similarity to their respective target molecule.

For each spectrum within a training batch, we chose a set of candidates given by the batch parameter, k_{aug}. Candidates selected for regularization in each batch are therefore the k_{aug} most similar candidates, and are taken sequentially in each training epoch. Figure 5.3 demonstrates the batching process for computing the regularization and contrastive losses. We then explored when and how to incorporate the regularization loss with our contrastive loss. Since our final ranking predictions are made using the molecular-spectral similarity in the joint space, the regularization attempts to push the most similar candidates away from the target molecule by minimizing a regularization loss function in addition to the contrastive loss. The regularization loss function minimizes the cosine similarity between the most similar candidates and the associated spectra - and hence the associated target molecules. The regularization loss, $L_{regularization}$, is defined as:

$$L_{regularization} = \frac{1}{k}\sum_{n=1}^{k}\frac{1}{k_{aug}}\sum_{m=1}^{k_{aug}}(cosine\_sim(z^n_{spec}, z^m_{cand})) \tag{5.6}$$

The total training loss, L_{total}, is the sum of the two losses weighted by hyper-parameters α and β:

$$L_{total} = \alpha * L_{contrastive} + \beta * L_{regularization} \tag{5.7}$$

We explored values for the α and β parameters, and we observed that utilizing regularization as a fine tuning strategy towards the end of the training provided the best performance. Regularization was turned on for the last 3% of the training epochs. The weight given to regularization loss was 10% to ensure that the matching pairs are not pushed too far apart during regularization. Therefore, $\alpha = 1.0$, $\beta = 0.0$ for first 97% epochs and $\alpha = 0.9$, $\beta = 0.1$ for last 3% of epochs. JESTR was trained on NVIDIA A100 GPU with 40GB of graphics RAM and 256GB of CPU RAM. Adam [122] was used as the optimizer. We used grid search over the ranges of the tuned parameters. The values of the parameters that achieved the best performance were selected and used to train and test the model for all datasets (Table 5.1).

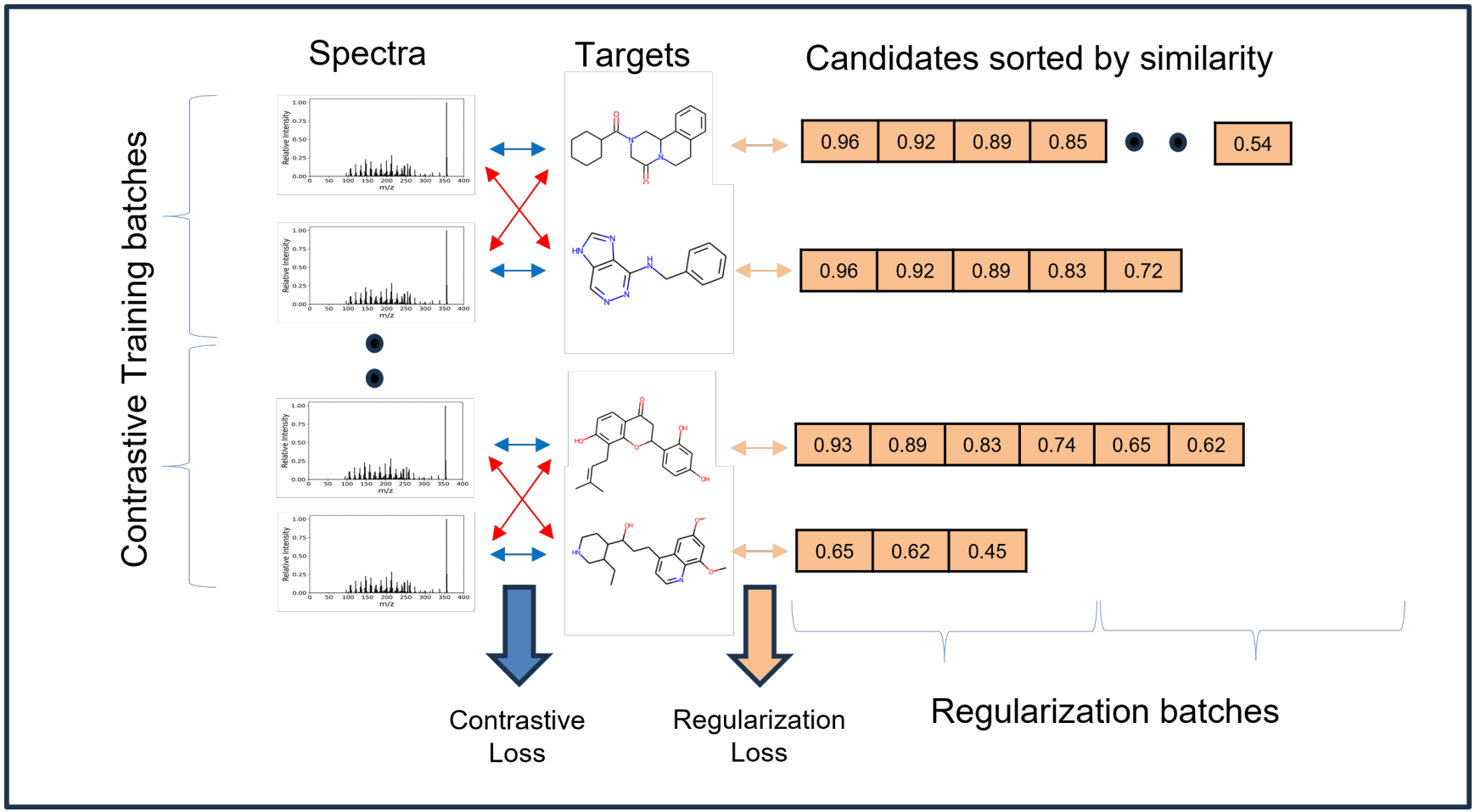


Figure 5.3 The contrastive loss is calculated over a batch of molecules at a time. For each molecule in the batch, candidates are selected from the candidate list sorted by similarity. If a particular molecule has a lesser number of candidates than the batch size, the candidates for that molecule will be repeated sequentially. The candidates thus selected are used to calculate the regularization loss. The curly braces in the vertical direction show the contrasting batching for two batches, while the curly braces in the horizontal direction show regularization batches for two batches

Table 5.1 Tuning of various hyperparameters for JESTR model using grid search.

Hyperparameter	Values searched	Final value
Contrastive Learning Rate	5e-6, 5e-5, 5e-4	5e-4
Contrastive Batch Size	32, 64, 128	32
Early Stopping Epochs	20, 40, 80	80
Regularization loss weight	0.1, 0.3, 0.5	0.1
% epochs used for regularization	3, 10, 20	3

We also looked at the training and inference times of all the models (Table 5.2). The runtimes were measured on a Linux machine with 48 CPU cores with 196GB RAM and 2 nVidia A100 GPUs with 40GB RAM. JESTR uses a GNN encoder for the molecules and an MLP encoder for the spectra. MIST uses MLP and transformers as its encoders for spectra, while ESP uses MLP and GNN as encoders for spectra and molecules, respectively. From a computational complexity of the model architecture, the three models are similar. The training time depends largely on the amount of training each model undergoes.

Table 5.2 Training time and inference time for the 3 models on the NPLIB1 dataset. The runtimes were measured on the same machine. JESTR takes longer to train because the training is run for a larger number of epochs.

Model	Training Time	Epochs	Inference Time
JESTR	6 hours	800	1.5 hours
MIST	1.2 hours	50	20 minutes
ESP	1.3 hours	100	1.2 hours

5.2 Experiment and results

We conduct experiments and analysis to answer the following research questions. (Q1) Does JESTR's implicit annotation method (with and without regularization) outperform prior explicit methods (mol-to-spec and spec-to-fp)? (Q2) Is learning the molecule-spectrum joint embedding space effective for distinguishing target molecules among their respective candidate sets? We compare JESTR against state-of-the-art mol-to-spec technique, ESP [95], and spec-to-FP technique, MIST [101]. We conduct the evaluation using three datasets:

the NPLIB1 dataset that was previously released with the CANOPUS tool [124], the well-curated, available-for-purchase NIST2020 dataset, and user-deposited data from MassBank of North America (MoNA) [88].

5.2.1 Datasets

The NPLIB1 dataset was first utilized by the CANOPUS tool to predict compound classes, e.g., benzenoids, phenol ethers, and others, from spectra, thus providing partial annotation on spectra in cases when spec-to-spec comparisons in reference spectral databases yield unsatisfactory matches [124]. This dataset was created by selecting spectra from the NIST2020 [87], GNPS [86] and MoNA [88] databases. The selection ensured a desired distribution of compound classes. This dataset was recently renamed to NPLIB1 [101] to distinguish it from the CANOPUS tool. We utilized the NPLIB1 data as assembled by MIST [101]. This dataset comprised 8,030 spectra measured under positive mode (positively charged, with an H adduct, [M+H]+) belonging to 7,131 unique target molecules. We utilize the same data split as proposed by MIST, where the split was structure-disjoint such that a molecule with the same InChiKeys did not appear both in the training and test sets. Therefore, 714 molecules and their 819 spectra were utilized for testing. Two additional datasets were utilized to explore training JESTR on larger datasets. The NIST2020 dataset is well-curated spectral database released by the National Institute of Standards and Technology. NIST2020 comprises a variety of molecules from human, bacteria, environmental, plant, and food samples. A variety of instruments and settings are used to measure spectra for each compound. The measurements are repeated and a consensus spectrum is created for each measurement. The NIST datasets are available under a commercial license, and we had access to the NIST2020 version of this dataset. The MassBank of North America (MoNA) is a collaborative database, with contributions by users. Both experimental and in-silico spectra are accepted. Here, we only retrieved the experimental dataset. Statistics for the three datasets is provided in Table 5.3. The number of unique molecules is largest in

NIST2020, while the number of spectra per molecule is the lowest in NPLIB1. The splits for the NIST2020 and MoNA were created ensuring that no molecules overlapped between the training and test sets.

Table 5.3 Spectra, molecule, and candidate statistics for the three datasets.

	Total		Train				Test			
Dataset	Spectra	Molecules	Spectra	Molecules	Max Cands	Avg Cands	Spectra	Molecules	Max Cands	Avg Cands
NPLIB1	8,030	7,131	7,211	6,417	44,374	2,220	819	714	25,929	2,274
NIST2020	291,515	22,001	262,408	19,800	42,542	1,390	29,107	2,201	42,376	1,322
MoNA	35,752	6,767	32,216	6,090	42,542	2,322	677	3,536	32,364	2,494

As typical in prior works [125, 126, 127], we select candidates for each target molecule in the training and test data by retrieving molecules from PubChem [128], by matching formulae of the target molecules. The average candidate sets for the target molecules range from 1,322 to 2,494 molecules, and for regularization, the average candidate sets for training molecules ranged from a 1,390 to 2,322 (Table 5.3).

5.2.2 Other annotation tools

To compare with other annotation models, we trained ESP [95] and MIST [101] on each of the datasets. The ESP model utilizes a GNN-based molecular encoder and an MLP on the molecular fingerprint. ESP is trained to learn a weighting between the molecular and fingerprint representations to predict the spectra. The best ESP performing model on rank@1 was the version that utilized the fingerprint and modeled peak co-dependencies, ESP MLP-PD. MIST first assigns chemical formulas to peaks within each spectrum using SIRIUS [84], and represents a spectrum as a set of chemical formulas. MIST trains a transformer model to learn peak embeddings and to predicts fingerprint. MIST also featurizes pairwise neutral losses and predicts substructure fragments as an auxiliary task. We ran both MIST and ESP on all three datasets, and confirmed the results with the respective teams.

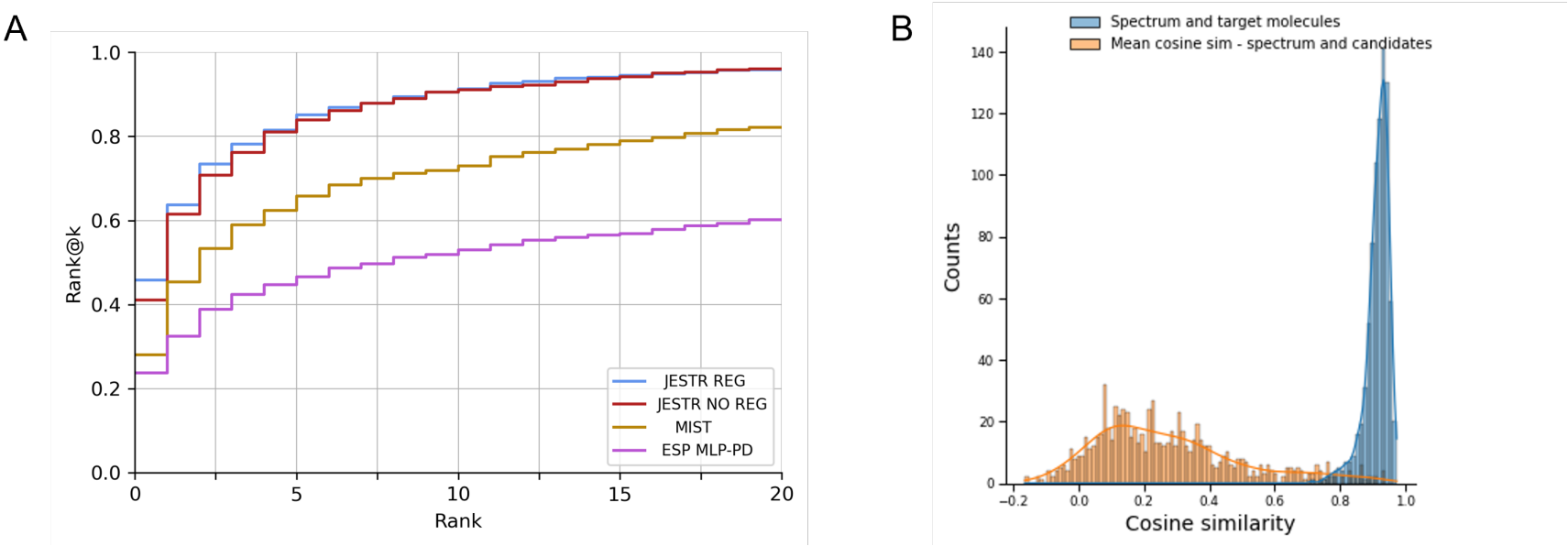


Figure 5.4 Regularization results for NPLIB1. A. Rank@k results for JESTR , with and without regularization, ESP MLP-PD, and MIST. B. Distribution of cosine similarities of query spectra and target/candidate molecules with contrastive learning using JESTR.

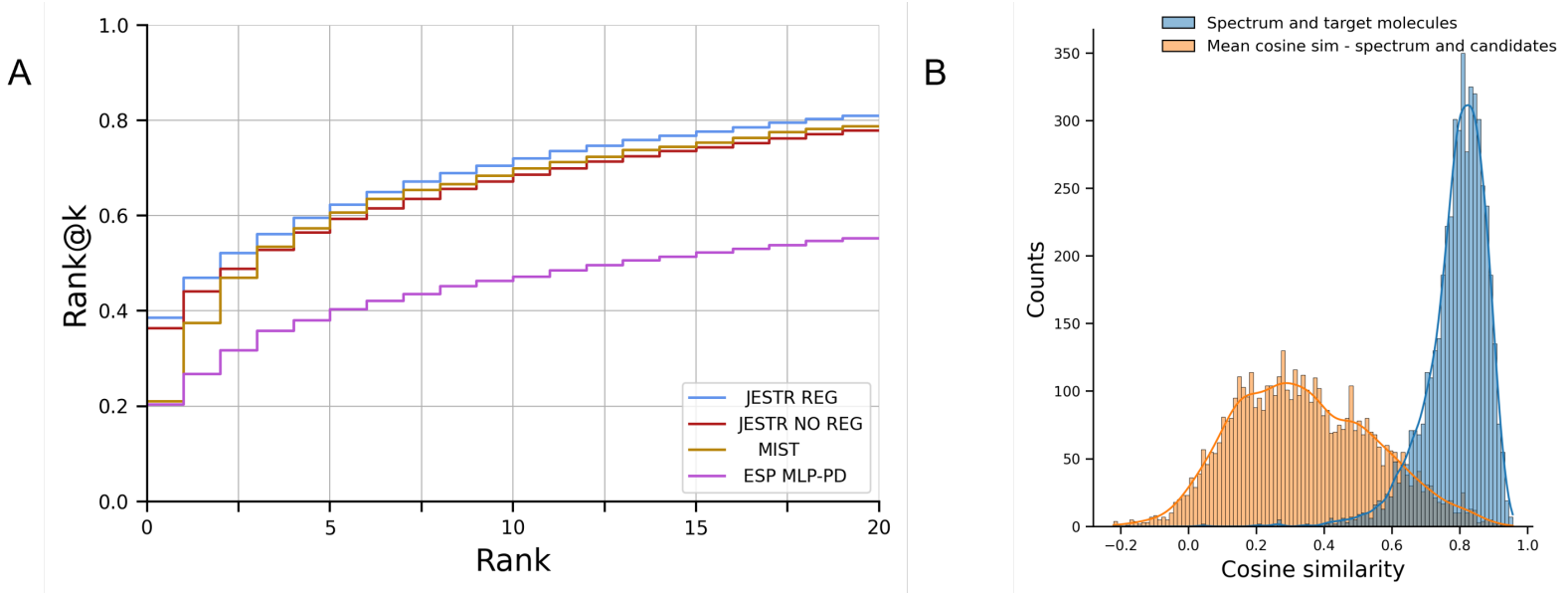


Figure 5.5 Regularization results for NIST2020. A. Rank@k results for JESTR on NIST2020, with and without regularization, MIST, and ESP MLP-PD. B. Distribution of cosine similarities of query spectra and target/candidate molecules in the NIST2020 test set with contrastive learning using JESTR.

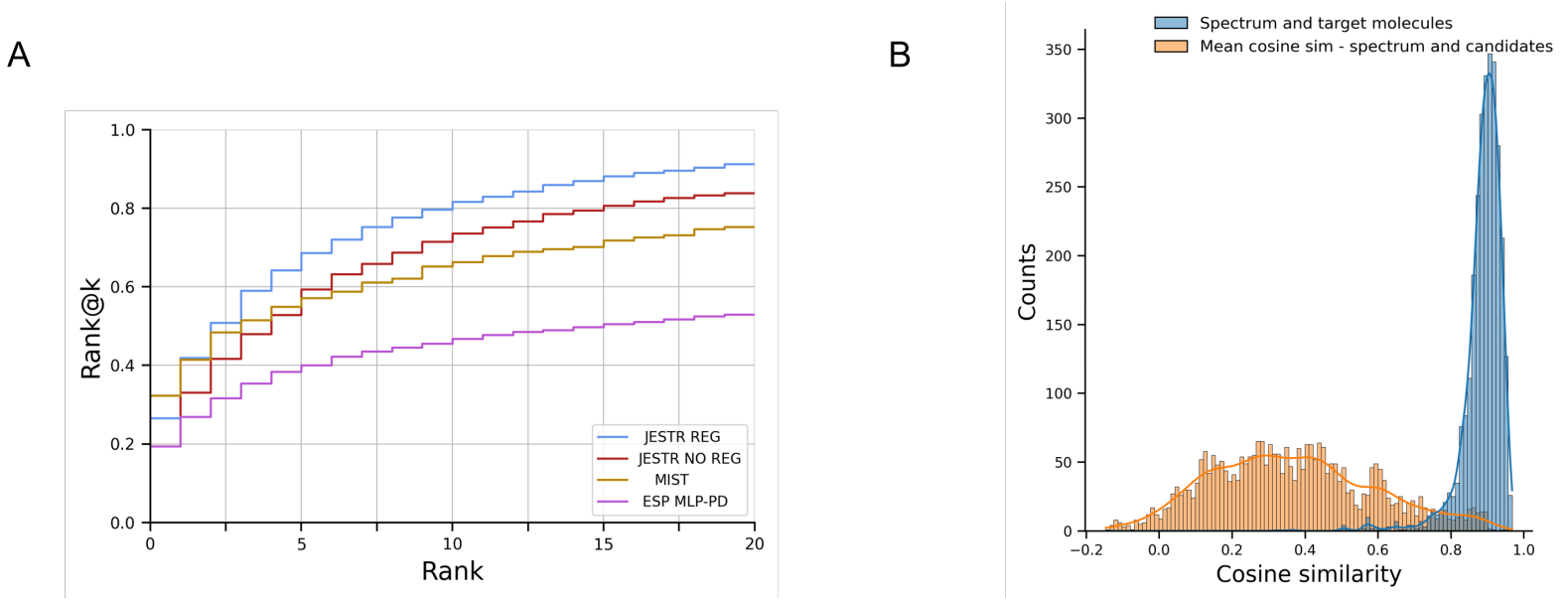


Figure 5.6 Regularization results for MoNA. A. Rank@k results for JESTR on MoNA, with and without regularization, MIST, and ESP MLP-PD. B. Distribution of cosine similarities of query spectra and target/candidate molecules in the MoNA test set with contrastive learning using JESTR.

Table 5.4 Ranking results for JESTR for the NPLIB1, NIST2020, and MoNA datasets. A. Ranking performance of all three tools B. Ablation study for the impact of removing regularization.

	NPLIB1			NIST2020			MoNA		
A. Comparison with other tools									
Model	Rank@1	Rank@5	Rank@20	Rank@1	Rank@5	Rank@20	Rank@1	Rank@5	Rank@20
ESP	23.7	44.7	60.2	20.5	30.1	48.8	19.4	38.4	53.0
MIST	27.9	62.4	82.1	21.0	57.4	78.8	**32.3**	52.9	75.3
JESTR (w/ regularization)	**45.8**	**81.5**	**95.8**	**38.6**	**59.6**	**81.0**	26.6	**64.2**	**91.2**
B. Ablation study									
JESTR (w/o regularization)	41.1	81.1	96.1	36.4	56.5	77.9	19.4	52.8	83.9

5.2.3 JESTR vs explicit-construction models

Given a query spectrum, the primary task of JESTR is to identify the target molecule among a set of candidates. Candidate ranking was therefore selected as the performance metric. The rank@1, rank@5 and rank@20 indicates the percentage of target molecules that were correctly ranked within the top 1, 5 and 20 candidates, respectively.

JESTR is compared against ESP and MIST (Table 5.4A). For the NPLIB1 dataset, JESTR outperforms ESP and MIST on all reported ranks. For rank@1, JESTR outperforms ESP by 93.2%, and MIST by 64.1%. Further, JESTR achieves 95.8% rank@20, while the maximum performance of ESP and MIST is at 60.2% and 82.1%, respectively. We plot the detailed ranks for the NPLIB1 dataset (Figure 5.4A). At all ranks, JESTR provides superior performance to both ESP and MIST. For the NIST2020 dataset, JESTR outperforms all other models. Specifically, JESTR outperforms MIST at rank@1 by 83.8% (Figure 5.5A). For the MoNA dataset, JESTR consistently outperforms ESP. MIST only outperforms JESTR at rank@1 by 17.6%, but not on rank@5, rank@20, or any other rank (Figure 5.6A). Across all datasets, on avarage for ranks@[1-5], JESTR outperforms ESP by 71.6% and MIST by 23.6%.

Examining the overall performance of JESTR over the three datasets, we note that JESTR's performance was worse on the MoNA dataset when compared to NPLIB1 and NIST2020. We suspect that JESTR's performance on MoNA was low for two reasons. First,

MoNA has the fewest number of molecules, thus providing the lowest molecular diversity among the three datasets. Second, MoNA data is uploaded by users and may not have undergone consistent curation efforts. The NIST2020 dataset is well curated; however, it has the highest ratio of spectra per molecule, an average of 13.25 spectra per molecule, versus 1.13 and 5.28 spectra per molecule for NPLIB1 and MoNA, respectively. As such, we suspect that JESTR finds it hard to place all the spectra embeddings closer to the molecule in the joint space for NIST2020 and MoNA. A combination of high molecule to spectra ratio, lower diversity of molecules and inconsistent spectra curation makes JESTR perform the lowest on MoNA. MIST, with additional information in the form of subformulae annotation on peaks, does a better job at distinguishing among various spectra of the same molecule for MoNA, thereby attaining a better rank@1 score on this dataset. However, MIST looses its advantage over JESTR starting with rank@2.

5.2.4 Joint-space embeddings distinguish target molecules from their candidates

The contrastive loss used in training JESTR ensures that the embeddings for matched spectrum-molecule pairs are placed close to each other in the joint embedding space, while non-matching spectrum-molecule pairs are placed further away. Figure 5.4B shows the distribution on spectrum-molecule cosine similarities for matching and non-matching pairs in the NPLIB1 test set. The corresponding distributions for NIST2020 and MoNA datasets are shown in Figure 5.5B and 5.6B respectively. While any molecule other than the target can be considered a non-matching partner for the query spectrum, Figure 5.4B only considers candidate molecules (same chemical formula as the target) as the non-matching partner. It is clear that JESTR well discriminates between target and candidate molecules.

5.2.5 Ablation study - removing regularization

To assess the value of regularization, the model was retrained without . Regularizing the training loss with molecular candidates improves rank@1 by 11.4%, 6.0% and 37.1% on the

NPLIB1, NIST2020 and MoNA datasets, respectively. Improvements using regularization is evident at almost all ranks and all datasets ((Table 5.4B). For rank@20 on NPLIB1, the results drop by 0.3% since the rank@20 result for NPLIB1 is high even without regularization, where even a small change in a few targets changes the result slightly.

To further demonstrate the value of regularization, we performed additional analysis on the NPLIB1 dataset. With regularization, the number of target molecules ranked @1 increases significantly, from 301 to 375, causing a ripple effect in improving other rank@k numbers (Figure 5.7A). Further, we examined the Tanimoto similarity between molecules in the training set and their candidates. We retrieved 15.86 million candidates from PubChem based on the chemical formulas of the target molecules in the training set. The majority of these candidates show low Tanimoto similarity with the target molecules (Figure 5.7B). Hence, our fine-tuning regularization strategy and sorting the candidates by their cosine similarity to the target effectively prioritizes regularization with the most similar candidates. We approximately utilized 7 million candidates for regularization during the last 3% of training epochs. Upon examining the cosine similarity distributions on the embeddings of candidate and target molecules (Figure 5.7C), we see that our regularization strategy reduces the average cosine similarity between targets and their candidates. Regularization is therefore effective, enabling the model to discriminate between a target molecule and its candidates. Similar analysis for the NIST2020 and MoNA datasets is shown in Figures 5.8 and 5.9.

5.3 Conclusion

JESTR offers a novel implicit annotation paradigm that avoids the explicit generation of spectra, fingerprints, or molecular structures. As molecules and spectra are views of the same object, embedding these views in a joint embedding space using contrastive learning provides a performance advantage. On NPLIB1, JESTR outperforms ESP by 88.9% and MIST by 41.1% on ranks@[1-5]. On NIST2020, JESTR outperforms ESP by 68.5% and

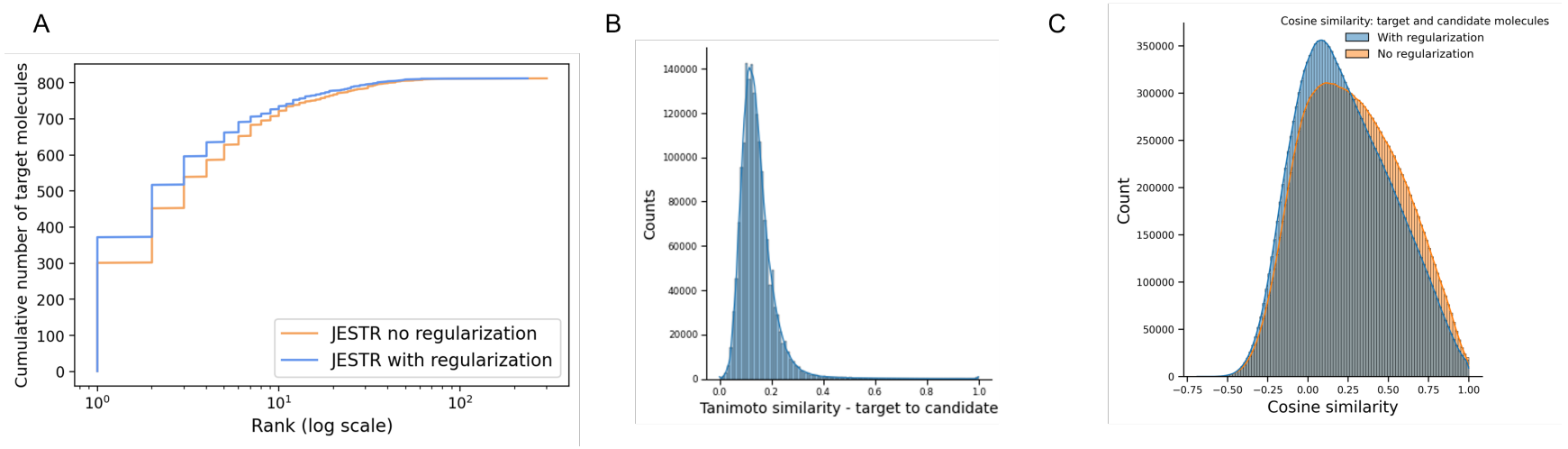


Figure 5.7 Regularization analysis for JESTR for NPLIB1. A. Regularization improves rank@k by significantly placing more targets at rank 1. B. Distribution on Tanimoto similarities on the ECFP fingerprints between target and candidates in the training set. C. Distribution on cosine similarities, with and without regularization, of the target and candidates within the test set.

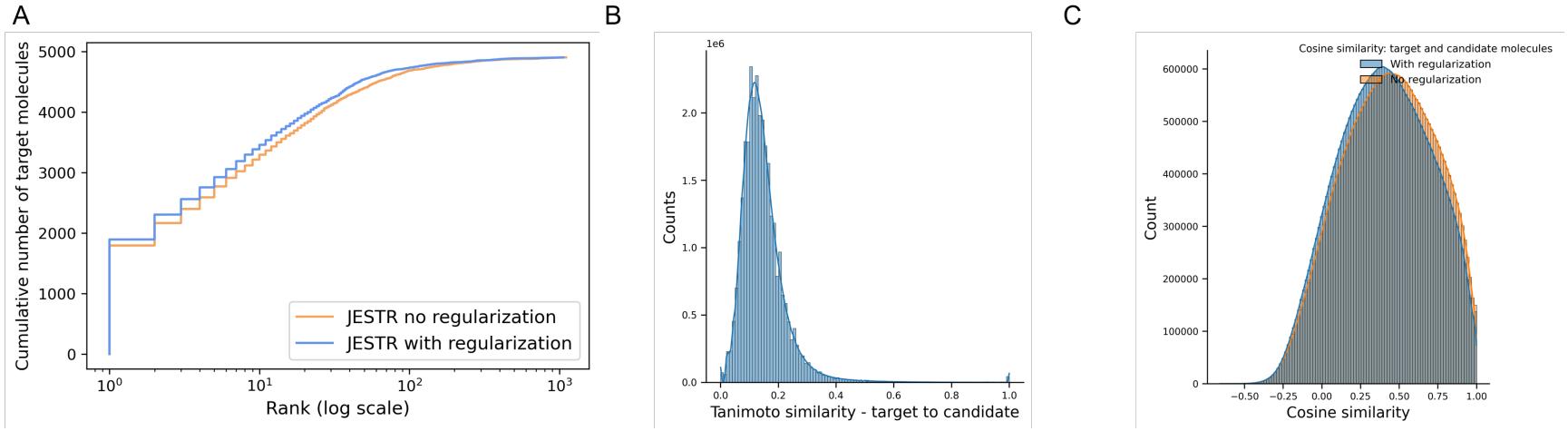


Figure 5.8 Regularization analysis for JESTR for NIST2020. A. Regularization improves rank@k by significantly placing more targets at rank 1. B. Distribution on Tanimoto similarities on the ECFP fingerprints between target and candidates in the training set. C. Distribution on cosine similarities, with and without regularization, of the target and candidates within the test set.

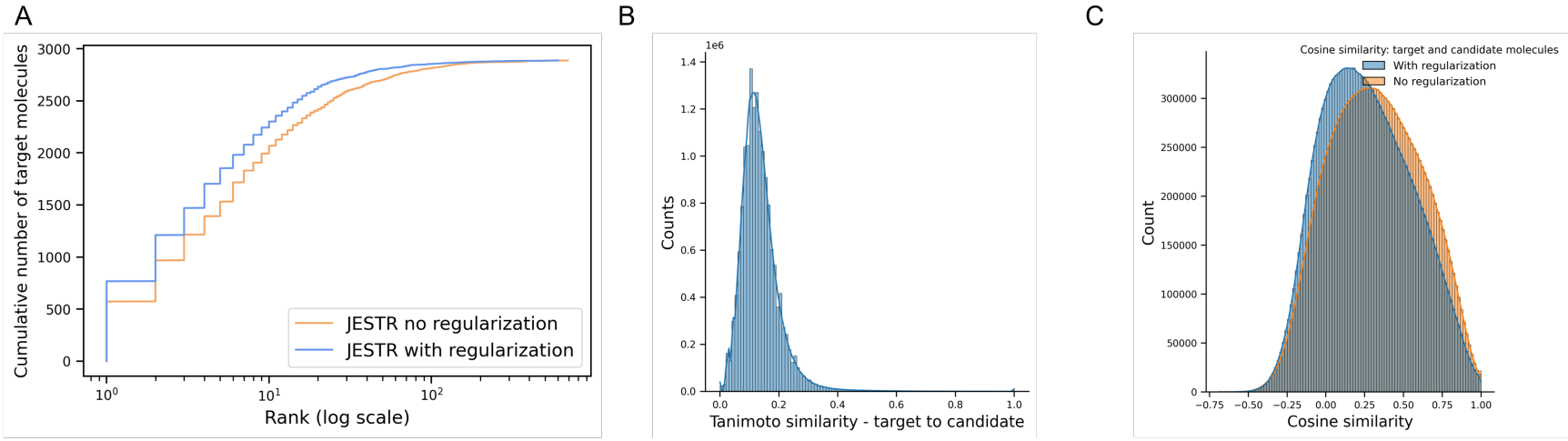


Figure 5.9 Regularization analysis for JESTR for MoNA. A. Regularization improves rank@k by significantly placing more targets at rank 1. B. Distribution on Tanimoto similarities on the ECFP fingerprints between target and candidates in the training set. C. Distribution on cosine similarities, with and without regularization, of the target and candidates within the test set.

MIST by 25.8% on average on rank@[1-5]. On MoNA, JESTR outperforms ESP by 57.5% and MIST by 4.0% on average on rank@[1-5]. Analysis of JESTR's performance on the three datasets reveals that dataset diversity, quality, and spectra-to-molecule ratios impact performance. Further, our results showed enhance value in utilizing candidate molecules during training for regularization, improving performance an average of 11.4% for all three datasets. The overall JESTR results are promising and vouch for the potential of implicit annotation approaches. We expect to attain further improvements by utilizing additional knowledge in the form of subformulae annotation on spectral peaks and by utilizing enhanced molecular and spectral encoders.

Chapter 6 - Conclusion and future research directions

This thesis has demonstrated the essential role of data quality and size in addressing three central challenges in Bioinformatics. First, for redox property prediction, we tackled the challenge of limited training data by assessing the contributions of molecular features, both individually and in various combinations. We trained a Gaussian Process Regression (GPR) model specifically optimized for small datasets, achieving robust predictions in data-limited scenarios. Second, in enzyme-molecule interaction prediction, we introduced an innovative method to stratify enzymatic reaction data, enabling contrastive learning to improve feature representations. This approach enhanced the quality of enzyme and molecule features, thereby enhancing interaction predictions. Finally, to facilitate the annotation of untargeted metabolomics spectra, we developed a method to embed molecular structures and spectra within a shared feature space. The shared embedding space allowed for effective ranking of candidate molecules based on similarity, enabling accurate annotation for complex spectra. Collectively, these advancements highlight the paramount importance of data quality and size in solving problems in Bioinformatics. Further, the work demonstrates that tailored machine learning techniques are indispensable for overcoming the unique challenges of noisy, sparse, or abundant biological datasets and advancing data-driven biological research.

6.1 Research summary

The first problem addressed in this thesis is redox potential prediction for organic molecules, a property that defines molecular behavior in numerous biological and biochemical applications. Multiple molecular features are evaluated, and a novel combination of molecular descriptors, ADME properties, and DFT energy-based features was selected for training a Gaussian Process Regression (GPR) model for each dataset. The model was trained and validated on three datasets and evaluated on relevant experimental dataset with measured redox values. Results demonstrate that careful selection of molecular features, in conjunction with a GPR-based model, delivers strong predictive performance in data-limited contexts.

The second problem focuses on predicting the likelihood of compound-enzyme interactions, an important step in understanding enzyme promiscuity. To tackle this problem, the Contrastive Stratification for Interaction Prediction (CSI) method was developed, introducing a novel data stratification approach that maximizes mutual information across multiple views of compound-enzyme interaction data. High-quality embeddings for molecules and enzyme sequences are learned through contrastive loss and are subsequently used to train a predictor for interaction likelihood. Evaluated on three diverse molecule-enzyme reaction datasets, CSI effectively enhances representation learning, demonstrating its capability to address data-abundant scenarios with increased predictive accuracy.

The final problem addressed in this thesis is annotation of spectral data, a challenge in interpreting metabolomics data collected from biological samples. The thesis introduces JESTR, a novel paradigm for annotating untargeted spectra. Traditional methods have generated intermediate representations, such as spectra or fingerprints, but remain limited in resolving untargeted spectra. JESTR innovates by embedding molecules and spectra in a shared embedding space, applying contrastive learning to bring congruent pairs close in the embedding space. Framing annotation as a ranking task, JESTR ranks candidate molecules based on the cosine similarity of their embeddings to that of the query spectrum .

When evaluated against other tools on three metabolomics datasets, JESTR almost always outperforms state of the art tools, underscoring the power of learning a joint embedding space in data-rich scenarios.

6.2 Future directions

In this thesis, we introduced novel approaches to Bioinformatics challenges in both data-limited and data-rich contexts. Our methods achieved state-of-the-art performance on test datasets, highlighting the potential of these approaches. There are several avenues for refinement and exploration that could further extend these results and open new possibilities for improved outcomes.

In the redox prediction project, we demonstrated how combinations of various molecular features contribute to the prediction of redox potential. A promising direction for future work involves leveraging pre-trained models to provide baseline molecular features. Several such models are available, trained on extensive molecular datasets such as cheminformatics repositories [129] and property databases. Tuning these models on redox potential data, we can potentially capture relevant chemical information more accurately. To further improve model interpretability and predictive power, integrating feature-weighting techniques like attention mechanisms could highlight the most informative molecular descriptors, refining the model's focus on redox-relevant features. Additionally, a phased screening approach could improve both the model's efficiency and scalability. Such an approach involves first generating predictions using low-cost computational descriptors, then applying more computationally intensive calculations to a filtered subset of promising molecules, thereby balancing speed and accuracy. Finally, training the model on a larger set of experimental redox potential data across diverse conditions and solvents would enhance its generalizability, leading to improved predictions across broader chemical spaces. Together, these enhancements would advance the model's applicability in redox property prediction.

In the enzyme-molecule interaction prediction project, our contrastive learning approach

using multiple data views showed promise, but several enhancements could further improve model performance. First, incorporating additional views based on binding pocket characteristics would allow the model to capture detailed features that drive enzyme specificity and catalytic efficiency. By stratifying data around binding pocket properties, we can highlight structural nuances that contribute to enzyme promiscuity and target selectivity. Another improvement involves leveraging pre-trained models of enzymes and molecules, especially those built on large-scale datasets, to generate initial embeddings of interacting components. Using these foundational representations could improve the robustness of enzyme and molecule features, potentially boosting prediction accuracy. Additionally, using harder negative samples, such as known enzyme inhibitors, could refine the model's ability to distinguish closely related interactions, leading to better prediction accuracy for true interactions. To further increase interpretability and focus, incorporating cross-attention between the molecular graph and amino acid residues within the enzyme's active site would allow the model to prioritize binding-relevant regions in both entities. Finally, extending the model beyond binary interaction prediction could broaden its utility. Instead of simply predicting interaction likelihood, the model could be trained to predict specific interaction properties, such as the binding pocket location, enzyme turnover rate (K_{cat}), and affinity via the Michaelis constant (K_m). Predicting interaction properties would yield valuable insights into enzyme functionality, advancing applications in drug discovery and metabolic engineering.

In the third project, the accuracy of the JESTR method can be further improved by enhancing the representations for both spectra and molecular structures. An initial direction would be annotating peaks with subformulae, which could provide more detailed information about the molecular structure, thus enriching the spectral data and its associated fragmentation patterns. Additionally, generating fragmentation trees for all potential fragmentation paths could provide more granular insights into molecular fragmentation, refining molecular representations by reflecting the diversity of possible fragmentation patterns. Another area

for improvement involves applying contrastive learning at a finer level, such as focusing on sub-spectrum and substructure data. Learning these finer-grained relationships would allow the model to better link specific molecular fragments with corresponding spectral features, increasing the accuracy of molecule-spectrum matching. Increasing the granularity of the joint embedding space through such strategies would facilitate more precise metabolite identification, improving the model's utility in identifying compounds from complex spectra. Furthermore, the model could establish connections between peaks and substructures, creating representations that incorporate structural context directly within the spectra and molecule embeddings. These advanced representations could then support the generation of unknown or novel metabolites from identified substructures, enhancing the utility of metabolomics, particularly in biomedical research where novel metabolites play key roles in biomarker discovery and disease diagnosis.

References

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

[2] A. O'Connell *et al.*, "Biocatalysis: Landmark discoveries and applications in chemical synthesis," *Chemical Society Reviews*, 2024.

[3] N.-J. Hu, S.-Y. Li, and Y.-C. Liu, "Recent advances in biocatalysis and metabolic engineering," *Catalysts*, vol. 11, no. 9, p. 1052, 2021.

[4] Y. Lu and N. M. Marshall, "Redox potential," in *Encyclopedia of Biophysics*, G. C. K. Roberts, Ed. Berlin, Heidelberg: Springer Berlin Heidelberg, 2013, pp. 2207–2211, ISBN: 978-3-642-16712-6.

[5] C. G. Chen, A. N. Nardi, A. Amadei, and M. D'Abramo, "Theoretical modeling of redox potentials of biomolecules," *Molecules*, vol. 27, no. 3, p. 1077, 2022.

[6] Y. Fu, L. Liu, H.-Z. Yu, Y.-M. Wang, and Q.-X. Guo, "Quantum-chemical predictions of absolute standard redox potentials of diverse organic molecules and free radicals in acetonitrile," *Journal of the American Chemical Society*, vol. 127, no. 19, pp. 7227–7234, 2005.

[7] N. S. Chandel, "Nadph—the forgotten reducing equivalent," *Cold Spring Harbor Perspectives in Biology*, vol. 13, no. 6, a040550, 2021.

[8] J. Vašková, L. Kočan, L. Vaško, and P. Perjési, "Glutathione-related enzymes and proteins: A review," *Molecules*, vol. 28, no. 3, p. 1447, 2023.

[9] T. Ikeda and K. Kano, "An electrochemical approach to the studies of biological redox reactions and their applications to biosensors, bioreactors, and biofuel cells," *Journal of bioscience and bioengineering*, vol. 92, no. 1, pp. 9–18, 2001.

[10] S. Li, C. D. G. Tavares, J. G. Tolar, and C. M. Ajo-Franklin, "Selective bioelectronic sensing of pharmacologically relevant quinones using extracellular electron transfer in lactiplantibacillus plantarum," *Biosensors and Bioelectronics*, vol. 243, p. 115 762, 2024.

[11] L. J. Bird *et al.*, "Engineering wired life: Synthetic biology for electroactive bacteria," *ACS Synthetic Biology*, vol. 10, no. 11, pp. 2808–2823, 2021.

[12] P. Bollella and E. Katz, "Enzyme-based biosensors: Tackling electron transfer issues," *Sensors*, vol. 20, no. 12, p. 3517, 2020.

[13] J. Zhang, F. Li, D. Liu, Q. Liu, and H. Song, "Engineering extracellular electron transfer pathways of electroactive microorganisms by synthetic biology for energy and chemicals production," *Chemical Society Reviews*, 2024.

[14] H. Chen *et al.*, "Fundamentals, applications, and future directions of bioelectrocatalysis," *Chemical Reviews*, vol. 120, no. 23, pp. 12 903–12 993, 2020.

[15] C. E. Pereyra, R. F. Dantas, S. B. Ferreira, L. P. Gomes, and F. P. Silva-Jr, "The diverse mechanisms and anticancer potential of naphthoquinones," *Cancer Cell International*, vol. 19, pp. 1–20, 2019.

[16] J. L. Bolton and T. Dunlap, "Formation and biological targets of quinones: Cytotoxic versus cytoprotective effects," *Chemical research in toxicology*, vol. 30, no. 1, pp. 13–37, 2017.

[17] R. Francke and R. D. Little, "Redox catalysis in organic electrosynthesis: Basic principles and recent developments," *Chemical Society Reviews*, vol. 43, no. 8, pp. 2492–2521, 2014.

[18] J. Liu, L. Lu, D. Wood, and S. Lin, "New redox strategies in organic synthesis by means of electrochemistry and photochemistry," *ACS Central Science*, vol. 6, no. 8, pp. 1317–1340, 2020.

[19] H. Tian, Z. Yu, A. Hagfeldt, L. Kloo, and L. Sun, "Organic redox couples and organic counter electrode for efficient organic dye-sensitized solar cells," *Journal of the American Chemical Society*, vol. 133, pp. 9413–22, Jun. 2011.

[20] P. T. Kissinger and W. R. Heineman, "Cyclic voltammetry," *Journal of chemical education*, vol. 60, no. 9, p. 702, 1983.

[21] D. S. Wishart, "Emerging applications of metabolomics in drug discovery and precision medicine," *Nature reviews Drug discovery*, vol. 15, no. 7, pp. 473–484, 2016.

[22] R. C. Prince, P. L. Dutton, and M. Gunner, "The aprotic electrochemistry of quinones," *Biochimica et Biophysica Acta (BBA)-Bioenergetics*, vol. 1863, no. 6, p. 148 558, 2022.

[23] E. Hruska, A. Gale, and F. Liu, "Bridging the experiment-calculation divide: Machine learning corrections to redox potential calculations in implicit and explicit solvent models," *Journal of Chemical Theory and Computation*, vol. 18, no. 2, pp. 1096–1108, 2022.

[24] Schrodinger, *Qikprop*, https://www.schrodinger.com/platform/products/qikprop/, 2024.

[25] D. J. Watson *et al.*, "Efficacies and adme properties of redox active methylene blue and phenoxazine analogues for use in new antimalarial triple drug combinations with amino-artemisinins," *Frontiers in Pharmacology*, vol. 14, p. 1 308 400, 2024.

[26] L. Tanner *et al.*, "¡i¿in vitro¡/i¿ efficacies, adme, and pharmacokinetic properties of phenoxazine derivatives active against mycobacterium tuberculosis," *Antimicrobial Agents and Chemotherapy*, vol. 63, no. 11, 10.1128/aac.01010–19, 2019. eprint: https://journals.asm.org/doi/pdf/10.1128/aac.01010-19.

[27] H.-J. Xue, X. Dai, J. Zhang, S. Huang, and J. Chen, "Deep matrix factorization models for recommender systems.," in *IJCAI*, Melbourne, Australia, vol. 17, 2017, pp. 3203–3209.

[28] X. He, L. Liao, H. Zhang, L. Nie, X. Hu, and T.-S. Chua, "Neural collaborative filtering," in *Proceedings of the 26th international conference on world wide web*, 2017, pp. 173–182.

[29] J. Vamathevan *et al.*, "Applications of machine learning in drug discovery and development," *Nature reviews Drug discovery*, vol. 18, no. 6, pp. 463–477, 2019.

[30] T. Yao, Y. Pan, Y. Li, and T. Mei, "Exploring visual relationship for image captioning," in *Proceedings of the European conference on computer vision (ECCV)*, 2018, pp. 684–699.

[31] Y. Bengio, A. Courville, and P. Vincent, "Representation learning: A review and new perspectives," *IEEE transactions on pattern analysis and machine intelligence*, vol. 35, no. 8, pp. 1798–1828, 2013.

[32] G. E. Hinton, A. Krizhevsky, and S. D. Wang, "Transforming auto-encoders," in *International conference on artificial neural networks*, Springer, 2011, pp. 44–51.

[33] I. Goodfellow *et al.*, "Generative adversarial networks," *Communications of the ACM*, vol. 63, no. 11, pp. 139–144, 2020.

[34] Y. Li, M. Yang, and Z. Zhang, "A survey of multi-view representation learning," *IEEE transactions on knowledge and data engineering*, vol. 31, no. 10, pp. 1863–1883, 2018.

[35] A. Radford *et al.*, "Learning transferable visual models from natural language supervision," in *International Conference on Machine Learning*, PMLR, 2021, pp. 8748–8763.

[36] Y. Tian, D. Krishnan, and P. Isola, "Contrastive multiview coding," in *European conference on computer vision*, Springer, 2020, pp. 776–794.

[37] S. Chopra, R. Hadsell, and Y. LeCun, "Learning a similarity metric discriminatively, with application to face verification," in *2005 IEEE computer society conference on computer vision and pattern recognition (CVPR'05)*, IEEE, vol. 1, 2005, pp. 539–546.

[38] T. Chen, S. Kornblith, M. Norouzi, and G. Hinton, "A simple framework for contrastive learning of visual representations," in *International conference on machine learning*, PMLR, 2020, pp. 1597–1607.

[39] P. Khosla *et al.*, "Supervised contrastive learning," *Advances in neural information processing systems*, vol. 33, pp. 18 661–18 673, 2020.

[40] J.-B. Grill *et al.*, "Bootstrap your own latent-a new approach to self-supervised learning," *Advances in neural information processing systems*, vol. 33, pp. 21 271–21 284, 2020.

[41] M. Assran *et al.*, "Self-supervised learning from images with a joint-embedding predictive architecture," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 15 619–15 629.

[42] A. Bardes, J. Ponce, and Y. LeCun, "Mc-jepa: A joint-embedding predictive architecture for self-supervised learning of motion and content features," *arXiv preprint arXiv:2307.12698*, 2023.

[43] K. He, X. Chen, S. Xie, Y. Li, P. Dollár, and R. Girshick, "Masked autoencoders are scalable vision learners," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2022, pp. 16 000–16 009.

[44] Y. Tian and Y. Zhang, "A comprehensive survey on regularization strategies in machine learning," *Information Fusion*, vol. 80, pp. 146–166, 2022.

[45] B. T. Blackburn *et al.*, "Identifying key properties that drive redox mediator activity in lactiplantibacillus plantarum," 2024.

[46] M. Heinonen, H. Shen, N. Zamboni, and J. Rousu, "Metabolite identification and molecular fingerprint prediction through machine learning," *Bioinformatics*, vol. 28, no. 18, pp. 2333–2341, 2012.

[47] R. C. Morrison, "The extended koopmans' theorem and its exactness," *The Journal of chemical physics*, vol. 96, no. 5, pp. 3718–3722, 1992.

[48] P. Lykos and G. Pratt, "Discussion on the hartree-fock approximation," *Reviews of Modern Physics*, vol. 35, no. 3, p. 496, 1963.

[49] D. P. Chong, O. V. Gritsenko, and E. J. Baerends, "Interpretation of the kohn–sham orbital energies as approximate vertical ionization potentials," *The Journal of Chemical Physics*, vol. 116, no. 5, pp. 1760–1772, 2002.

[50] J. Stewart, "Optimization of parameters for semiempirical methods vi: More modifications to the nddo approximations and re-optimization of parameters," *Journal of molecular modeling*, vol. 19, Nov. 2012.

[51] T. Bredow and K. Jug, "Theory and range of modern semiempirical molecular orbital methods," *Theoretical Chemistry Accounts*, vol. 113, pp. 1–14, 2005.

[52] R. Fedorov and G. Gryn'ova, "Unlocking the potential: Predicting redox behavior of organic molecules, from linear fits to neural networks," *Journal of Chemical Theory and Computation*, vol. 19, no. 15, pp. 4796–4814, 2023.

[53] D. Rogers and M. Hahn, "Extended-connectivity fingerprints," *Journal of chemical information and modeling*, vol. 50, no. 5, pp. 742–754, 2010.

[54] G. Landrum, *Getting started with the rdkit in python—the rdkit 2020.03. 1 documentation*, 2016.

[55] S. Ghule, S. R. Dash, S. Bagchi, K. Joshi, and K. Vanka, "Predicting the redox potentials of phenazine derivatives using dft-assisted machine learning," *ACS omega*, vol. 7, no. 14, pp. 11 742–11 755, 2022.

[56] N. Schneider, D. M. Lowe, R. A. Sayle, and G. A. Landrum, "Development of a novel fingerprint for chemical reactions and its application to large-scale reaction classification and similarity," *Journal of chemical information and modeling*, vol. 55, no. 1, pp. 39–53, 2015.

[57] C.-H. Li and D. P. Tabor, "Discovery of lead low-potential radical candidates for organic radical polymer batteries with machine-learning-assisted virtual screening," *Journal of Materials Chemistry A*, vol. 10, no. 15, pp. 8273–8282, 2022.

[58] A. Jinich, B. Sanchez-Lengeling, H. Ren, R. Harman, and A. Aspuru-Guzik, "A mixed quantum chemistry/machine learning approach for the fast and accurate prediction of biochemical redox potentials and its large-scale application to 315 000 redox reactions," *ACS central science*, vol. 5, no. 7, pp. 1199–1210, 2019.

[59] J. Wang, "An intuitive tutorial to gaussian processes regression," *Computing in Science & Engineering*, 2023.

[60] X. Chen *et al.*, "Drug–target interaction prediction: Databases, web servers and computational models," *Briefings in bioinformatics*, vol. 17, no. 4, pp. 696–712, 2016.

[61] M. Bagherian, E. Sabeti, K. Wang, M. A. Sartor, Z. Nikolovska-Coleska, and K. Najarian, "Machine learning approaches and databases for prediction of drug–target interaction: A survey paper," *Briefings in bioinformatics*, vol. 22, no. 1, pp. 247–269, 2021.

[62] L. Zhou *et al.*, "Revealing drug-target interactions with computational models and algorithms," *Molecules*, vol. 24, no. 9, p. 1714, 2019.

[63] K. Abbasi, P. Razzaghi, A. Poso, S. Ghanbari-Ara, and A. Masoudi-Nejad, "Deep learning in drug target interaction prediction: Current and future perspectives," *Current Medicinal Chemistry*, vol. 28, no. 11, pp. 2100–2113, 2021.

[64] L. F. Salas-Nuñez *et al.*, "Machine learning to predict enzyme–substrate interactions in elucidation of synthesis pathways: A review," *Metabolites*, vol. 14, no. 3, p. 154, 2024.

[65] B.-X. Du *et al.*, "Compound–protein interaction prediction by deep learning: Databases, descriptors and models," *Drug discovery today*, vol. 27, no. 5, pp. 1350–1366, 2022.

[66] G. M. Visani, M. C. Hughes, and S. Hassoun, "Enzyme promiscuity prediction using hierarchy-informed multi-label classification," *Bioinformatics*, vol. 37, no. 14, pp. 2017–2024, 2021.

[67] Q. Feng, E. Dueva, A. Cherkasov, and M. Ester, "Padme: A deep learning-based framework for drug-target interaction prediction," *arXiv preprint arXiv:1807.09741*, 2018.

[68] X. Lin, "Deepgs: Deep representation learning of graphs and sequences for drug-target binding affinity prediction," *arXiv preprint arXiv:2003.13902*, 2020.

[69] I. Lee, J. Keum, and H. Nam, "DeepConv-DTI: Prediction of drug-target interactions via deep learning with convolution on protein sequences," *PLoS computational biology*, vol. 15, no. 6, e1007129, 2019.

[70] M. Tsubaki, K. Tomii, and J. Sese, "Compound–protein interaction prediction with end-to-end learning of neural networks for graphs and sequences," *Bioinformatics*, vol. 35, no. 2, pp. 309–318, 2019.

[71] T. Nguyen, H. Le, T. P. Quinn, T. Nguyen, T. D. Le, and S. Venkatesh, "GraphDTA: Predicting drug–target binding affinity with graph neural networks," *Bioinformatics*, vol. 37, no. 8, pp. 1140–1147, 2021.

[72] S. Min, S. Park, S. Kim, H.-S. Choi, B. Lee, and S. Yoon, "Pre-training of deep bidirectional protein sequence representations with structural information," *IEEE Access*, vol. 9, pp. 123 912–123 926, 2021.

[73] K. Huang, C. Xiao, L. M. Glass, and J. Sun, "MolTrans: Molecular Interaction Transformer for drug–target interaction prediction," *Bioinformatics*, vol. 37, no. 6, pp. 830–836, 2021.

[74] T. Wang *et al.*, "Deepenzyme: A robust deep learning model for improved enzyme turnover number prediction by utilizing features of protein 3d-structures," *Briefings in Bioinformatics*, vol. 25, no. 5, bbae409, 2024.

[75] J. Deng, Y. Zhang, Y. Pan, X. Li, and M. Lu, "Multidta: Drug-target binding affinity prediction via representation learning and graph convolutional neural networks," *International Journal of Machine Learning and Cybernetics*, pp. 1–10, 2024.

[76] S. Decherchi and A. Cavalli, "Thermodynamics and kinetics of drug-target binding by molecular simulation," *Chemical Reviews*, vol. 120, no. 23, pp. 12 788–12 833, 2020.

[77] M. K. Gilson, T. Liu, M. Baitaluk, G. Nicola, L. Hwang, and J. Chong, "BindingDB in 2015: A public database for medicinal chemistry, computational chemistry and systems pharmacology," *Nucleic acids research*, vol. 44, no. D1, pp. D1045–D1053, 2016.

[78] A. Chang *et al.*, "BRENDA, the ELIXIR core data resource in 2021: New developments and updates," *Nucleic Acids Research*, vol. 49, no. D1, pp. D498–D508, 2021.

[79] M. Kanehisa, M. Furumichi, Y. Sato, M. Ishiguro-Watanabe, and M. Tanabe, "KEGG: Integrating viruses and cellular organisms," *Nucleic acids research*, vol. 49, no. D1, pp. D545–D551, 2021.

[80] J. Tang *et al.*, "Making sense of large-scale kinase inhibitor bioactivity data sets: A comparative and integrative analysis," *Journal of Chemical Information and Modeling*, vol. 54, no. 3, pp. 735–743, 2014.

[81] M. I. Davis *et al.*, "Comprehensive analysis of kinase inhibitor selectivity," *Nature biotechnology*, vol. 29, no. 11, pp. 1046–1051, 2011.

[82] R. Schmid *et al.*, "Integrative analysis of multimodal mass spectrometry data in mzmine 3," *Nature biotechnology*, vol. 41, no. 4, pp. 447–449, 2023.

[83] S. Xing, S. Shen, B. Xu, X. Li, and T. Huan, "Buddy: Molecular formula discovery via bottom-up ms/ms interrogation," *Nature Methods*, vol. 20, no. 6, pp. 881–890, 2023.

[84] K. Dührkop *et al.*, "Sirius 4: A rapid tool for turning tandem mass spectra into metabolite structure information," *Nature methods*, vol. 16, no. 4, pp. 299–302, 2019.

[85] T. Kind *et al.*, "Identification of small molecules using accurate mass ms/ms search," *Mass spectrometry reviews*, vol. 37, no. 4, pp. 513–532, 2018.

[86] M. Wang *et al.*, "Sharing and community curation of mass spectrometry data with global natural products social molecular networking," *Nature biotechnology*, vol. 34, no. 8, pp. 828–837, 2016.

[87] NIST, *NIST20: Updates to the nist tandem and electron ionization spectral libraries*, https://www.nist.gov/programs-projects/tandem-mass-spectral-library/, 2020.

[88] MoNA, *MassBank of North America*, https://mona.fiehnlab.ucdavis.edu/, 2024.

[89] M. Martin, W. Bittremieux, and S. Hassoun, "Molecular structure discovery for untargeted metabolomics using biotransformation rules and global molecular networking," *bioRxiv*, pp. 2024–02, 2024.

[90] S. Wolf, S. Schmidt, M. Müller-Hannemann, and S. Neumann, "In silico fragmentation for computer assisted identification of metabolite mass spectra," *BMC bioinformatics*, vol. 11, pp. 1–12, 2010.

[91] C. Ruttkies, E. L. Schymanski, S. Wolf, J. Hollender, and S. Neumann, "Metfrag relaunched: Incorporating strategies beyond in silico fragmentation," *Journal of cheminformatics*, vol. 8, pp. 1–16, 2016.

[92] F. Wang, J. Liigand, S. Tian, D. Arndt, R. Greiner, and D. S. Wishart, "Cfm-id 4.0: More accurate esi-ms/ms spectral prediction and compound identification," *Analytical chemistry*, vol. 93, no. 34, pp. 11 692–11 700, 2021.

[93] J. N. Wei, D. Belanger, R. P. Adams, and D. Sculley, "Rapid prediction of electron–ionization mass spectrometry using neural networks," *ACS central science*, vol. 5, no. 4, pp. 700–708, 2019.

[94] H. Zhu, L. Liu, and S. Hassoun, "Using graph neural networks for mass spectrometry prediction," *arXiv preprint arXiv:2010.04661*, 2020.

[95] X. Li, Y. Zhou Chen, A. Kalia, H. Zhu, L.-p. Liu, and S. Hassoun, "An ensemble spectral prediction (ESP) model for metabolite annotation," *Bioinformatics*, vol. 40, no. 8, btae490, 2024.

[96] A. Young, B. Wang, and H. Röst, "Massformer: Tandem mass spectrum prediction for small molecules using graph transformers," *arXiv preprint arXiv:2111.04824*, 2021.

[97] NIH, *Pubchem*, https://pubchem.ncbi.nlm.nih.gov, 2024.

[98] M. A. Stravs, K. Dührkop, S. Böcker, and N. Zamboni, "Msnovelist: De novo structure generation from mass spectra," *Nature Methods*, vol. 19, no. 7, pp. 865–870, 2022.

[99] E. E. Litsa, V. Chenthamarakshan, P. Das, and L. E. Kavraki, "An end-to-end deep learning framework for translating mass spectra to de-novo molecules," *Communications Chemistry*, vol. 6, no. 1, p. 132, 2023.

[100] T. Butler *et al.*, "Ms2mol: A transformer model for illuminating dark chemical space from mass spectra," *ChemRxiv. 2023; doi:10.26434/chemrxiv-2023-vsmpx-v3*, 2023.

[101] S. Goldman, J. Wohlwend, M. Stražar, G. Haroush, R. J. Xavier, and C. W. Coley, "Annotating metabolite mass spectra with domain-inspired chemical formula transformers," *Nature Machine Intelligence*, vol. 5, no. 9, pp. 965–979, 2023.

[102] H. L. Morgan, "The generation of a unique machine description for chemical structures-a technique developed at chemical abstracts service.," *Journal of chemical documentation*, vol. 5, no. 2, pp. 107–113, 1965.

[103] D. Weininger, "Smiles, a chemical language and information system. 1. introduction to methodology and encoding rules," *Journal of chemical information and computer sciences*, vol. 28, no. 1, pp. 31–36, 1988.

[104] scikit, *Scaler*, 2020.

[105] H. Neugebauer, F. Bohle, M. Bursch, A. Hansen, and S. Grimme, "Benchmark study of electrochemical redox potentials calculated with semiempirical and dft methods," *The Journal of Physical Chemistry A*, vol. 124, no. 35, pp. 7166–7176, 2020.

[106] M. T. Huynh, C. W. Anson, A. C. Cavell, S. S. Stahl, and S. Hammes-Schiffer, "Quinone 1 e–and 2 e–/2 h+ reduction potentials: Identification and analysis of deviations from systematic scaling relationships," *Journal of the American Chemical Society*, vol. 138, no. 49, pp. 15 903–15 910, 2016.

[107] J. Tirado-Rives and W. L. Jorgensen, "Performance of b3lyp density functional methods for a large set of organic molecules," *Journal of chemical theory and computation*, vol. 4, no. 2, pp. 297–306, 2008.

[108] O. Buriez and E. Labbé, "Disclosing the redox metabolism of drugs: The essential role of electrochemistry," *Current Opinion in Electrochemistry*, vol. 24, pp. 63–68, 2020.

[109] W. J. Egan, "Predicting adme properties in drug discovery," *Drug design: structure- and ligand-based approaches*, pp. 165–177, 2010.

[110] Schrodinger, *Ligprep*, https://www.schrodinger.com/platform/products/ligprep/, 2024.

[111] Monolith, *Gaussian process regression*, https://support.monolithai.com/support/solutions/articles/80001072565-gaussian-processes-regression, 2023.

[112] C. E. Rasmussen, "Gaussian processes in machine learning," in *Summer school on machine learning*, Springer, 2003, pp. 63–71.

[113] J. Gasteiger and M. Marsili, "Iterative partial equalization of orbital electronegativity—a rapid access to atomic charges," *Tetrahedron*, vol. 36, no. 22, pp. 3219–3228, 1980.

[114] Y. Tian, C. Sun, B. Poole, D. Krishnan, C. Schmid, and P. Isola, "What makes for good views for contrastive learning?" *Advances in Neural Information Processing Systems*, vol. 33, pp. 6827–6839, 2020.

[115] A. Tripathi and V. A. Bankaitis, "Molecular docking: From lock and key to combination lock," *Journal of molecular medicine and clinical applications*, vol. 2, no. 1, 2017.

[116] D. J. Lipman and W. R. Pearson, "Rapid and sensitive protein similarity searches," *Science*, vol. 227, no. 4693, pp. 1435–1441, 1985.

[117] G. Landrum, "Rdkit documentation," *Release*, vol. 1, no. 1-79, p. 4, 2013.

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

[119] H. Öztürk, A. Özgür, and E. Ozkirimli, "DeepDTA: Deep drug–target binding affinity prediction," *Bioinformatics*, vol. 34, no. 17, pp. i821–i829, 2018.

[120] T. He, "SimBoost: A Read-Across Approach for Drug-Target Interaction Prediction Using Gradient Boosting Machines," Ph.D. dissertation, Applied Sciences: School of Computing Science, 2016.

[121] A. Cichonska *et al.*, "Computational-experimental approach to drug-target interaction mapping: A case study on kinase inhibitors," *PLoS computational biology*, vol. 13, no. 8, e1005678, 2017.

[122] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.

[123] M. Kotera, S. Goto, and M. Kanehisa, "Predictive genomic and metabolomic analysis for the standardization of enzyme data," *Perspectives in Science*, vol. 1, no. 1-6, pp. 24–32, 2014.

[124] K. Dührkop *et al.*, "Systematic classification of unknown metabolites using high-resolution fragmentation mass spectra," *Nature biotechnology*, vol. 39, no. 4, pp. 462–471, 2021.

[125] K. Dührkop, H. Shen, M. Meusel, J. Rousu, and S. Böcker, "Searching molecular structure databases with tandem mass spectra using csi: Fingerid," *Proceedings of the National Academy of Sciences*, vol. 112, no. 41, pp. 12 580–12 585, 2015.

[126] F. Huber, S. van der Burg, J. J. van der Hooft, and L. Ridder, "Ms2deepscore: A novel deep learning similarity measure to compare tandem mass spectra," *Journal of cheminformatics*, vol. 13, no. 1, p. 84, 2021.

[127] N. F. de Jonge *et al.*, "Ms2query: Reliable and scalable ms2 mass spectra-based analogue search," *Nature Communications*, vol. 14, no. 1, pp. 1–12, 2023.

[128] S. Kim *et al.*, "Pubchem 2019 update: Improved access to chemical data," *Nucleic acids research*, vol. 47, no. D1, pp. D1102–D1109, 2019.

[129] S. Chithrananda, G. Grand, and B. Ramsundar, "Chemberta: Large-scale self-supervised pretraining for molecular property prediction," *arXiv preprint arXiv:2010.09885*, 2020.