Text2Mol: Cross-Modal Molecule Retrieval with Natural Language Queries

Carl Edwards, ChengXiang Zhai, Heng Ji
University of Illinois Urbana-Champaign
{cne2, czhai, hengji}@illinois.edu

Abstract

We propose a new task, **Text2Mol**, to retrieve molecules using natural language descriptions as queries. Natural language and molecules encode information in very different ways, which leads to the exciting but challenging problem of integrating these two very different modalities. Although some work has been done on text-based retrieval and structure-based retrieval, this new task requires integrating molecules and natural language more directly. Moreover, this can be viewed as an especially challenging cross-lingual retrieval problem by considering the molecules as a language with a very unique grammar. We construct a paired dataset of molecules and their corresponding text descriptions, which we use to learn an aligned common semantic embedding space for retrieval. We extend this to create a cross-modal attention-based model for explainability and reranking by interpreting the attentions as association rules. We also employ an ensemble approach to integrate our different architectures, which significantly improves results from 0.372 to 0.499 MRR. This new multimodal approach opens a new perspective on solving problems in chemistry literature understanding and molecular machine learning.[1]

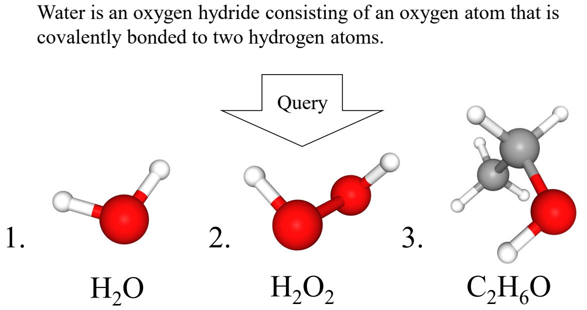

Water is an oxygen hydride consisting of an oxygen atom that is covalently bonded to two hydrogen atoms.
Query
1.
2.
3.
H_2O
H_2O_2
C_2H_6O

Figure 1: Given a natural language description of water, we want to rank the corresponding molecule H_2O first among all the possible molecules.

1 Introduction

Discovering new properties and applications of different molecules is critical for accelerating discovery in medicine and science. Existing databases contain tens of millions of molecules; PubChem (Kim et al., 2016, 2019) alone has 110 million compounds. Many information retrieval (IR) tools for chemistry rely on queries based on natural language descriptions of the molecules and existing chemical reactions. Hundreds of millions of possible molecules cannot all possibly undergo laboratory experimentation and be given attention by experts in order to create a description. To address this issue, it is critical to retrieve molecules directly from natural language descriptions. This approach allows newly discovered molecules to be easily integrated into the proposed IR framework. Our framework also allows for semantic-level search between natural language descriptions and molecules as well as for query expansion within traditional chemistry information retrieval systems.

Over the past several years, chemists have begun to rely increasingly on computational techniques for cataloging molecules and predicting chemical reactions, products, and properties, such as yield, toxicity, and water solubility (Wu et al., 2018; Glavatskikh et al., 2019; Coley et al., 2017; Ahneman et al., 2018; Fooshee et al., 2018). However, natural language and molecules are very different modalities of data, which makes integrating them together a challenging task. We argue that these two modalities are complementary and should be considered together.

Much current work focuses on images and language (Mogadala et al., 2020), but it is beneficial for the community to consider modalities beyond traditional ones, increasing their work's impact and efficacy. For example, integrating NLP and

[1] The programs and data are publicly available at github.com/cnedwards/text2mol for research purposes.

Molecule	An electrically neutral group of atoms bonded together.
Compound	Two or more elements held together by chemical bonds.
Chemical fingerprint	Represents a molecule or substructure using a bitstring. This allows for efficient substructure search and similarity calculation.
Morgan fingerprint	A specific type of chemical fingerprint also known as ECFP.
SMILES string	A character-based sequence representation of a molecule. (for example, C1=CC=CC=C1 is the SMILES string for benzene)
Canonical SMILES	A unique SMILES string for a molecule.

Table 1: Relevant Terminology

molecules could improve drug discovery and design.

In pursuit of this goal, we propose a multimodal embedding approach for constructing an aligned semantic space between these two types of data to allow for cross-modal retrieval. No previous work has studied this retrieval problem. The closest is (Zhou et al., 2010), which uses a hybrid approach to document retrieval by replacing chemicals in text with canonical keywords in order to standardize different chemical synonyms. However, this does not take the semantic information of the molecule (properties beyond the atoms and graph structure, such as being a pollutant or hydrophobic) into account.

Additionally, incorporating cross-modal attention can lead to insights on the relations between molecule substructures and text keywords. For example, we find that given "pollutant," the model focuses on the substructure $F-C$. This contributes to higher-level explainability between molecules and their descriptions.

Our molecular encoder is based on the Mol2vec (Jaeger et al., 2018) algorithm, which creates "sentences" of substructure identifiers from molecules; we frame **Text2Mol** as a new, particularly challenging type of cross-lingual information retrieval (CLIR). This problem is much more challenging than traditional CLIR since the gap between the query and target is much larger. It also provides a useful benchmark for extending CLIR to incorporate multiple data modalities. Molecules are essentially a different language with a uniquely challenging grammar. In fact, several techniques apply models developed for natural language processing to SMILES strings—machine-readable character-based representations for molecules (Weininger, 1988; Weininger et al., 1989).

The major novel contributions of this paper are:

- A new task **Text2Mol**: Cross-modal Text-Molecule Information Retrieval directly from natural language descriptions to molecules.
- Cross-modal attention-based association rules between molecules and text are used to improve results and for explainability.
- A new benchmark dataset with 33,010 text-compound pairs for cross-modal text-molecule IR which can be used for cross-lingual, multimodal, and explainable IR.

2 Task Definition

To push the boundaries of multimodal models, we present a new IR task: **Text2Mol**.

Given a text query and list of molecules without any reference textual information (represented, for example, as SMILES strings, graphs, or other equivalent representations) retrieve the molecule corresponding to the query. Figure 1 shows an example of this task. From a text description of a molecule, the model must incorporate the information in the description into a semantic representation which can be used to directly retrieve the molecule.

This requires the integration of two very different types of information: the structured knowledge represented by text and the chemical properties present in molecular graphs. We assume there is only one correct (relevant) molecule for each description, so we consider two measures for this task: Hits@1 and mean reciprocal rank (MRR).

3 Related Work

3.1 Multimedia Representation

Much recent work in this area has fallen into the category of vision-language models which leverage transformers (Chen et al., 2019; Su et al., 2020; Lu et al., 2019). There are also more fine-grained multimedia embedding approaches, such as integrating events from images and their descriptions (Li et al., 2020) or multimodal pattern mining (Li et al., 2016). CLIP (Radford et al., 2021) uses natural language to train a zero-shot image classifier which can be easily applied to different datasets. Specifically, their loss function, which follows Sohn (2016), serves as a very efficient version of binary cross-entropy loss by comparing all samples in a mini-batch with each other. To our knowledge, we are the first to apply this technique to molecules and text, and we also extend this loss function to incorporate negative samples to allow for cross-modal attention between the two encoders.

3.2 Molecule Representation

One critical problem in the field of molecular machine learning is molecule representation. Fingerprinting methods have long been employed in cheminformatics to featurize molecule structural representations (Cereto-Massagué et al., 2015; Sandfort et al., 2020). However, this approach does not allow these representations to be learned from the data. Other representations include techniques such as kernel PCA using Tanimoto similarity (Rensi and Altman, 2017; Mallory et al.). Recent advances in machine learning have begun to be applied to this problem. Jaeger et al. (2018) use the Morgan fingerprinting algorithm to convert each molecule into a 'sentence' of its substructures. A dataset of molecules can be interpreted as a corpus, and Mol2vec then applies Word2vec (Mikolov et al., 2013a,b) to create molecule representations. Additionally, other recent advances such as BERT (Devlin et al., 2019) have been applied to the domain such as MolBERT (Fabian et al., 2020) and ChemBERTa (Chithrananda et al., 2020), which use SMILES strings (Weininger et al., 1989) as inputs to pretrain a BERT-esque model.

3.3 Substructure or Description Retrieval

Although the biomedical domain has been more popular than chemistry (Zheng et al., 2014; Li et al., 2019; Li and Ji, 2019; Islamaj Doğan et al., 2019; Zhang et al., 2021; Lai et al., 2021), information retrieval in chemistry has long been studied and is summarized by Krallinger et al. (2017). Most work has focused on only a single modality: text or molecules. Text-based retrieval includes tasks such as finding relevant papers for a chemical or reaction and chemical entity recognition. Much work has also been done in graph and molecule-based retrieval (Hagadone, 1992; Barnard, 1993; Yan et al., 2005; Kratochvíl et al., 2018; Qu et al., 2019; Kratochvíl, 2019; Goyal et al., 2020). Hybrid approaches have also been attempted; Zhou et al. (2010) replace chemical entities in text with a unique canonical key (thus standardizing synonyms). This also allows them to perform query expansion by including similar molecules from their database. In contrast to this, we perform direct semantic cross-modal retrieval task in our approach, as opposed to just augmenting queries. Work in chemical entity recognition has also incorporated hybrid approaches, mostly as chemical name to structure converters such as ChemSpot (Rocktäschel et al., 2013) and OPSIN (Lowe et al., 2011).

3.4 Cross-Lingual Retrieval

Cross-lingual information retrieval (CLIR) is a technique to retrieve documents from a target language given a query in a different source language. Two common strategies are either translating the query into the target language or translating the document corpus into the source language (Zhang and Zhao, 2020). Further, work exists combining these approaches using interlingual semantics, such as via bilingual word embeddings (Vulic and Moens, 2015) or word embeddings and a dictionary (Bhattacharya et al., 2016).

Our problem, cross-modal molecule retrieval from text, can be considered as a CLIR task which we approach using an interlingual semantic approach. The model is trained on a parallel corpus of molecules and descriptions.

4 Methodology

4.1 Model

To accomplish this retrieval task, we need to connect text to molecules. To do so, we build an aligned semantic embedding space. Our approach consists of two distinct submodels: a text encoder and a molecule encoder. Both submodels create an embedding in the aligned space, and cosine similarity is used to rank the embeddings. A description embedding can be compared against a database of existing molecule embeddings, and this process scales easily using an approximate nearest neighbor search algorithm such as (Johnson et al., 2017). For the text encoder, we use SciBERT (Beltagy et al., 2019) and a linear projection to the embedding space followed by layer normalization (Ba

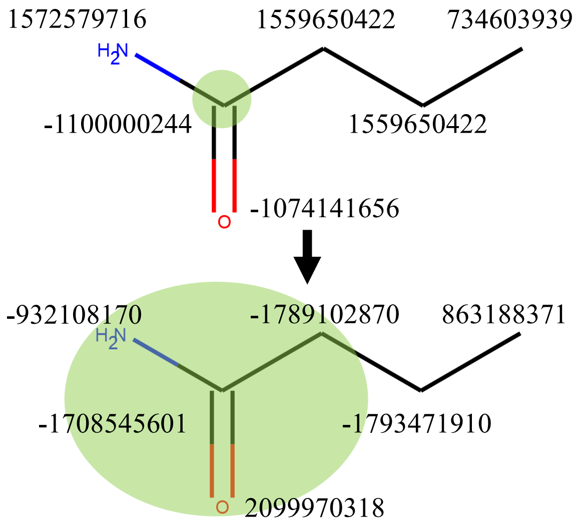


Figure 2: Example of Morgan Fingerprinting from (Rogers and Hahn, 2010) for Butyramide. The algorithm updates the identifiers from radius $r = 0$ to $r = 1$, as shown by the green circles.

et al., 2016). For the molecule encoder, we consider two architectures. First, we use a multi-layer perceptron (MLP) that takes Mol2vec embeddings as input. Second, we integrate a graph convolutional network (GCN) (Kipf and Welling, 2017) into Mol2vec.

Mol2vec (Jaeger et al., 2018) converts molecule graph structures into "sentences" of substructures. These substructures are created using Morgan fingerprinting (Rogers and Hahn, 2010), which is a type of topological fingerprint, which were historically used for quick substructure lookup. Morgan fingerprints incorporate a number of molecular properties based on the Daylight atomic invariants rule (Weininger et al., 1989). Atomic invariants such as the number of connections, number of non-hydrogen bonds, and atomic number are used to create the initial identifier for an atom. By using a circular hashing technique, they are able to create a unique identifier for a molecular substructure of some radius r centered around a central atom, as shown in Figure 2. The algorithm starts with a radius of zero which is iteratively increased until the desired substructure size is obtained. In Mol2vec, these fingerprints are used as tokens for each atom. In this work, we use a default value of $r = 1$, which gives two tokens for each atom ($r = 0$ and $r = 1$). This set of tokens is canonicalized in the same way as the canonical SMILES representation (Weininger et al., 1989). This list of tokens can be interpreted as a "sentence", and Mol2vec builds a corpus of such sentences. It then uses the Word2vec skip-gram (Mikolov et al., 2013a,b) algorithm to create "word" embeddings, which it averages together to create molecule representations. We use a two-layer MLP followed by a linear projection and layer normalization to create a trainable representation from the Mol2vec embedding, followed by layer normalization.

While the Morgan fingerprints (substructure tokens) incorporate some implicit graph information, we explicitly introduce the molecule graph structure using a GCN that takes a molecular graph as input with Mol2vec token embeddings as features. For example, rings are very important substructures in molecules. If the description mentions "aromatic ring" or "phenyl group," we want to be able to match this substructure in the molecule. We could potentially do so by increasing the maximum radius of the Morgan fingerprinting algorithm, but then there might not be enough examples of the resulting large-radius tokens to create a good representation given our corpus size. Particularly for large molecules, to capture the global structural information, we might need a very large radius which will create a lot of rare tokens (that get replaced by the UNK token). Instead, we explicitly incorporate the graph structure by using a GCN.

The Mol2vec token features are input to a three-layer GCN to create node representations for each atom in the molecule. These representations are combined using global mean pooling, and passed through two more hidden layers to produce a molecule representation. Since Mol2vec produces multiple tokens based on Morgan fingerprints of different radii, we select the corresponding token with the largest radius.

4.2 Cross-Modal Attention Model

To improve the explainability of our approach, we introduce a model with cross-modal attention by modifying the base model to use a transformer decoder (Vaswani et al., 2017). This decoder uses the SciBERT output as a source sequence and the node representations from the Mol2vec GCN model as a target sequence, and the attentions can be used to learn multimodal association rules. The architecture is shown in Figure 3.

4.3 Loss

To optimize the models, we base our loss on the symmetric contrastive loss presented by Radford et al. (2021). The loss takes the output embed-

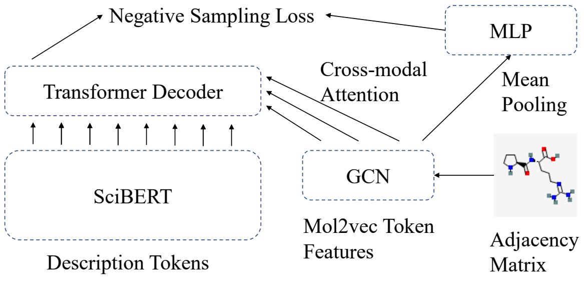


Figure 3: Model architecture for the cross-modal attention extension and association rules.

dings of both submodels, multiplies by the exponent of a learned temperature parameter, τ, and then takes the outer product of the mini-batch. The identity matrix I is used as labels. Categorical cross-entropy (CCE) is then applied along both axes, and the two losses are summed. This improves efficiency by allowing all the other samples in a mini-batch to serve as negatives. It corresponds to cosine similarity because the normalized dot product is minimized or maximized, for positives and negatives respectively. For batch embedding m and t of length n,

$$L(m,t) = CCE(e^{\tau} mt^{T}, I_n) + CCE(e^{\tau} tm^{T}, I_n)$$

We find this loss to be ineffective for training the cross-modal attention model because it encourages the model to ignore the textual information—i.e. information can leak from one encoder to the other. To remedy this problem, we modify this loss function to incorporate a matching task by introducing negative samples. We randomly sample new descriptions and replace their respective ones in the diagonal of the identity matrix with zeros, creating a binary classification task—does the description match the molecule? Since the rows with all zeros are no longer probability distributions, we instead use binary cross-entropy loss. This modified loss provides more signal than a pure matching task since it also receives signal from the other negatives, and it enforces the constraint that the model consider both the molecule and text description.

4.4 Cross-Modal Reranking

We want to better understand how the base networks work, so we introduce a modified model with cross-modal attention, which we also use to rerank the output of the base models. Given a training set of molecule-text pairs, P, we first train the cross-modal matching model. We collect the attention weights of the final layer for each of these pairs. Next, the attention weights for molecule token, m, and text token, t, are aggregated to create association rules. We define the support for a rule r from t to m as the sum of all attentions,

$$supp(r) = \sum_{p \in P} \sum_{\substack{t' \in p_t \\ m' \in p_m}} \mathbb{1}_{\substack{t=t' \\ m=m'}} a_{t',m'}$$

where $a_{i,j}$ is the attention weight between tokens i and j and p_t and p_m are the multisets of text and molecule tokens in p, respectively.

This produces association rules from every text token t to every molecule token m. We calculate the confidence for each of these rules by taking the support of the rule and dividing by the support of all the rules using t,

$$conf(t \implies m) = \frac{supp(t,m)}{\sum_{t' \in T} supp(t',m)}$$

where T is the set of all text tokens.

Following this, given a molecule and text pair, we consider all association rules that can produce it, and we take the average of the top k confidence values. For association-rule based reranking, Bharadwaj et al. (2014) takes the average of all confidence values. However, they have a comparatively smaller number of confidence rules. On the other hand, AnyBURL (Meilicke et al., 2019, 2020) finds maximum aggregation to be most effective. It also shows rule-based approaches can be very efficient (Ott et al., 2021). For our approach, we want to consider multiple one-to-one rules because we only use rules from one text token to one molecule token since the computational costs scale significantly due to the combinatorial number of many-to-many rules. By taking an average of only the top confidence values, we incorporate multiple one-to-one rules but ignore unimportant rules. This combines the two approaches to reranking while keeping in mind efficiency. We calculate a score by interpolating between the cosine similarities with association rule-based scores (AR) linearly,

$$S(a,b) = \alpha \cos(a,b) + (1-\alpha)AR(a,b)$$

where $\alpha \in [0,1]$ is selected on the validation set.

4.5 Ensemble Approach

Upon investigation of the baseline models, we found that the correct molecule was very frequently found in the top molecules. However, many of the molecules ranked above the correct molecule did not occur in the top results of the same model trained with different parameter initialization. We found that by taking an average of these rankings, the correct molecule's average rank would stay roughly the same, but the average rank of false positives increases. When these average ranks are used to reorder the results, the order of the incorrect and correct molecule switches. We find this method to be surprisingly effective, and we connect this to committee of neural networks (Drucker et al., 1994) in ensemble learning (Polikar, 2012). Additionally, we draw comparisons to Mixture of Experts-based models (Masoudnia and Ebrahimpour, 2014) such as Fan et al. (2006) and the Switch Transformer which contains 1.6 trillion parameters (Fedus et al., 2021). We compute the score, S, as a weighted average,

$$S(m) = \sum_i w_i R_i(m) \qquad s.t. \sum_i w_i = 1$$

for some molecule m where R_i is the rank assigned to that molecule by model i and w_i is the model weight. A lower score is more desirable.

5 Experiments

5.1 Data and Evaluation

For our task, we create a dataset using PubChem (Kim et al., 2016, 2019) and Chemical Entities of Biological Interest (ChEBI) (Hastings et al., 2016). We collect ChEBI annotations of compounds scraped from PubChem, which consists of 102,980 compound-description pairs. Using this data, we create a dataset consisting of 33,010 pairs, which we call **ChEBI-20**, that contains descriptions of more than 20 words. We find that longer descriptions tend to be less noisy and more informative. We remove compounds which cannot be processed by RDKit (Landrum, 2021).

We separate these datasets into 80%/10%/10% train-validation-test splits. The alignment models are trained on the training data, and the results are evaluated by searching all molecules in the dataset. The molecules in the training set are processed by Mol2vec using default parameters: a radius of 1, a threshold for unknown tokens of 3, an embedding dimension of 300, and a window size of 10.

5.2 Results

To train the models, we use Adam optimizer (Kingma and Ba, 2015) and two different learning rates. The SciBERT model uses a finetuning learning rate of 3e-5, as used by Devlin et al. (2019). The rest of the model uses 1e-4 as used by Vaswani et al. (2017). We use a linear annealing rate for the learning rate with 1,000 steps of warmup. We train for 40 epochs with a batch size of 32. We also use a temperature value of 0.07 as suggested by Radford et al. (2021). We use the first 256 text tokens for the text encoder.

5.2.1 Baseline Models

The MLP and GCN encoder models both show similar performance. Three results for both are shown in Table 2. We believe the performance similarity between MLP and GCN is because the description is a bottleneck. However, they appear to be effective at different tasks. In the test set, the mean rank is significantly lower for the MLP models than the GCN models; however, the MRR values are fairly similar. This indicates that these two architectures have different strengths. Further, the difference in mean rank is much smaller in the validation set; the validation mean rank is 30.60 and 28.89 for the MLP and GCN ensembles respectively. This indicates that the GCN architecture is more effective for retrieving the most difficult examples in the validation set (since there are not outlier ranks to increase the mean), but the MLP is more effective at difficult examples in the test set. We further examine this in Section 5.4.

5.2.2 Ensemble

We find that the ensemble method shows significant performance improvements. The ensemble of the three GCN models increases Test Hits@1 by roughly 8% from the baseline models. It is notable that the hyperparameters for these models are exactly the same, and the models are learning different ways of ranking which are complementary. To combine the different models, we find the heuristic of using uniform weights to be very effective.

A further advantage of the ensemble approach is that it can incorporate different encoder architectures and retrieval schemes, which may have different understandings of how to solve the problem. We find that combining both architectures is

	Training				Test			
Model	Mean Rank	MRR	Hits@1	Hits@10	Mean Rank	MRR	Hits@1	Hits@10
MLP1	9.55	0.428	26.5%	77.5%	30.38	0.372	22.4%	68.6%
MLP2	9.82	0.425	26.4%	77.1%	30.72	0.369	22.3%	68.9%
MLP3	9.53	0.431	26.9%	77.8%	36.30	0.372	22.3%	67.9%
GCN1	10.22	0.432	27.2%	76.5%	42.28	0.366	21.7%	68.2%
GCN2	9.67	0.423	26.7%	77.4%	41.90	0.371	22.3%	68.9%
GCN3	10.12	0.420	25.8%	76.7%	39.11	0.366	22.3%	67.9%
MLP-Ensemble	5.81	0.520	35.1%	86.4%	20.78	0.452	29.4%	77.6%
GCN-Ensemble	6.09	0.516	35.0%	86.1%	28.77	0.447	29.4%	77.1%
All-Ensemble	**4.67**	**0.568**	**40.2%**	**89.8%**	**20.21**	**0.499**	**34.4%**	**81.1%**
MLP1+Attn					30.37	0.375	22.8%	68.7%
MLP1+FPGrowth					30.37	0.374	22.6%	68.6%

Table 2: Results. FPGrowth is the frequent pattern growth algorithm (Han and Pei, 2000). Models 1, 2, and 3 only differ in initial parameter initialization.

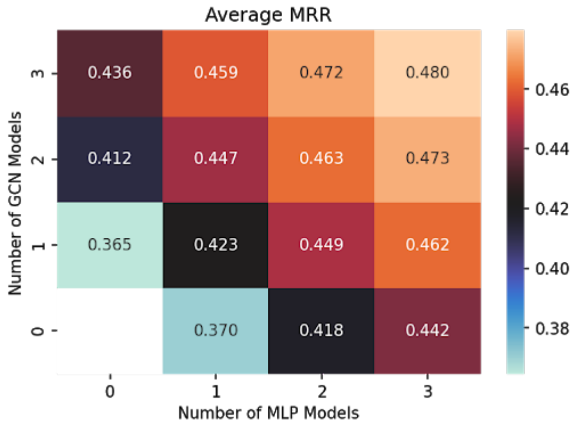


Figure 4: Validation MRR values for different combinations of architectures. The axes indicate the number of each architecture used. Ensembles with both architectures are more effective.

more effective than either alone; this is shown in Figure 4. Ensembles that only incorporate one architecture are consistently outperformed by models that incorporate both. For example, using 3 MLP models has an MRR of 0.442 but using 2 MLP and 1 GCN model has an MRR of 0.449.

5.3 Cross-Modal Attention and Reranking

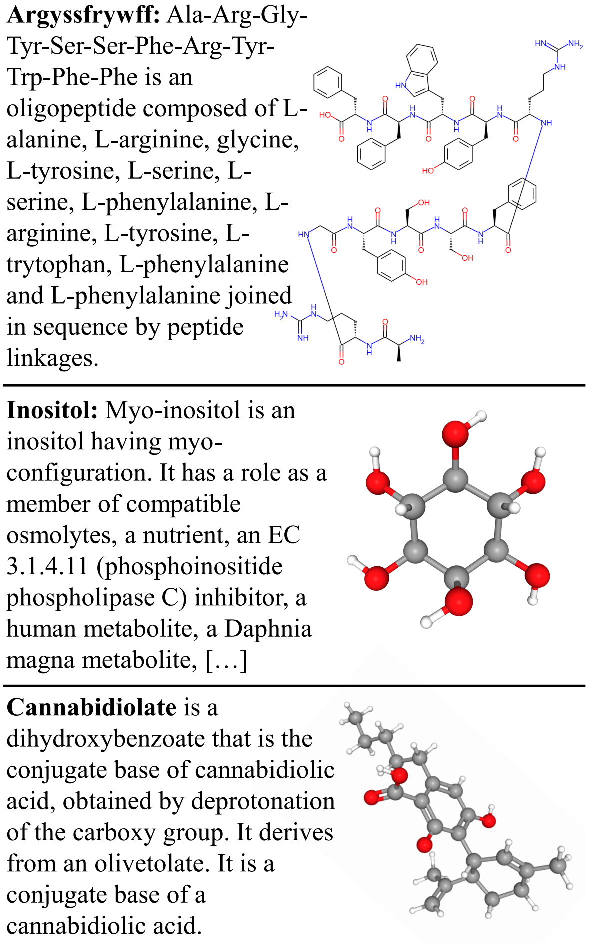
To better understand the behavior of the model, we apply cross-modal attention using a transformer decoder with 3 layers, and we rerank the top 10 of MLP1 using the 10 most confident association rules. We find cross-modal reranking to slightly improve our baseline model and to outperform traditional association rule mining, which can be accomplished by the FPGrowth algorithm (Han and Pei, 2000). Hits@1 for the baseline MLP model is increased by about 0.4%, but normal association rules only improve it by 0.2%.

Token	Substructure	Supp	Conf
Titanium	$Ti=O$	1.29	0.65
Aluminium	Al^{3+}	4.31	0.23
Manganese	Mn^{2+}	10.08	0.30
Toluene	$C-C=C$	12.93	0.231
Toluene	C_7H_8	23.79	0.425
##chloro	$Cl-C$	18.81	0.207
pollutant	$F-C$	3.097	0.208
chromatography	$C-Si$	2.976	0.271
acid	$C-O-H$	2398.7	0.078
crown	$C-C-O$	4.18	0.325

Table 3: Examples of interesting learned association rules from token to substructure. C_7H_8 is the chemical formula of toluene.

Mining these rules using attention also allows us to understand the connections the model is making. Examples of these rules are shown in Table 3. We primarily examine one-to-one rules; however, these one-to-one rules will often "split" the confidence among themselves. For example, toluene is a ring containing different substructures, so there will be multiple one-to-one rules required to capture the substructure. The rule from toluene to the three common substructure tokens in toluene has an increased confidence and support. Since we average the confidence values of all applicable rules, this is accounted for in reranking.

One interesting phenomenon we find is that the model is very interested in O-H structures (hydroxyl groups). It is also interested in positively charged metal ions in salts. The token "acid" has many different rules; however, the most confident



Argyssfrywff: Ala-Arg-Gly-Tyr-Ser-Ser-Phe-Arg-Tyr-Trp-Phe-Phe is an oligopeptide composed of L-alanine, L-arginine, glycine, L-tyrosine, L-serine, L-serine, L-phenylalanine, L-arginine, L-tyrosine, L-trytophan, L-phenylalanine and L-phenylalanine joined in sequence by peptide linkages.

Inositol: Myo-inositol is an inositol having myo-configuration. It has a role as a member of compatible osmolytes, a nutrient, an EC 3.1.4.11 (phosphoinositide phospholipase C) inhibitor, a human metabolite, a Daphnia magna metabolite, [. . .]

Cannabidiolate is a dihydroxybenzoate that is the conjugate base of cannabidiolic acid, obtained by deprotonation of the carboxy group. It derives from an olivetolate. It is a conjugate base of a cannabidiolic acid.

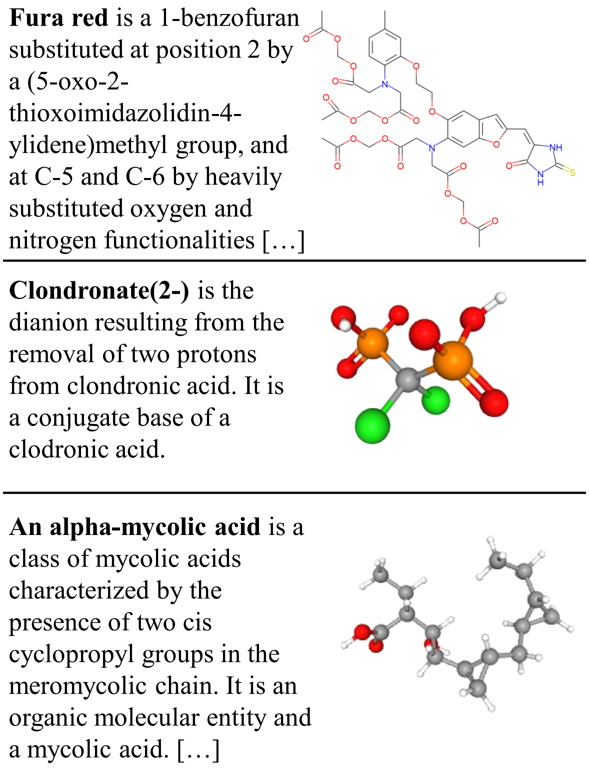
Figure 5: Example queries that are predicted correctly by All-Ensemble.

is a hydroxyl (-OH) group, which matches basic chemical properties of acids. Rules involving rare tokens can result in high confidence values. For example, the rule "crown" implies $C-C-O$ has a confidence of 0.325. This is because the dataset contains two "crown ether" molecules which have multiple occurrences of $C-C-O$.

5.4 Qualitative Analysis

Our technique is capable of retrieving large, complicated molecules as well as small ones. For example, it successfully retrieves both Argyssfrywff ($C_{79}H_{99}N_{19}O_{17}$) and Inositol ($C_6H_{12}O_6$), shown in Figure 5. Argyssfrywff shows that the model is capable of composing molecules from constituent parts mentioned in the description.

The MLP and GCN models capture different aspects of the molecules leading to different rankings. For example, MLP-ensemble ranks an alpha-mycolic acid ($C_{15}H_{26}O_3$) at 43; GCN-ensemble ranks it 3. The compound contains cyclopropyl groups (the triangles), shown in Figure 6, which the GCN captures. On the other hand, Clondronate(2-) ($CH_2Cl_2O_6P_2^{-2}$) is ranked 4,915 by the GCN but 61 by the MLP, showing large differences exist between the architectures. The models are also mutually beneficial; 2-Methylideneglutaric acid ($C_6H_8O_4$) is ranked 2nd by MLP and 3rd by GCN, but it is ranked 1st by All-Ensemble. Individual models trained identically (but with different initial parameters) also show this phenomenon. GCN 1, 2, and 3 rank Pierreione C ($C_{27}H_{28}O_6$) 2nd. GCN1 ranks Aspernidine A 1st, but it is ranked 49 and 64 by GCN 2 and 3, respectively. The average rank of Aspernidine A becomes 38, so GCN-Ensemble ranks Pierreione C 1st.



Fura red is a 1-benzofuran substituted at position 2 by a (5-oxo-2-thioxoimidazolidin-4-ylidene)methyl group, and at C-5 and C-6 by heavily substituted oxygen and nitrogen functionalities [. . .]

Clondronate(2-) is the dianion resulting from the removal of two protons from clondronic acid. It is a conjugate base of a clodronic acid.

An alpha-mycolic acid is a class of mycolic acids characterized by the presence of two cis cyclopropyl groups in the meromycolic chain. It is an organic molecular entity and a mycolic acid. [. . .]

Figure 6: Example queries that are ranked incorrectly by All-Ensemble.

The model is able to ignore irrelevant description information. For example, MLP achieves rank 1 for Rostratin D ($C_{18}H_{20}N_2O_6S_4$), whose description includes the unique and likely unuseful section "isolated from the whole broth of the marine-derived fungus Exserohilum rostratum." Instead, the model successfully identifies it from the following attributes: "bridged compound, a cyclic ketone, a lactam, an organic disulfide, an organic heterohexacyclic compound, a secondary alcohol, a dithiol and a diol."

There are some very challenging queries where multiple molecules are very similar. For example, Pro-Arg and Arg-Pro share the same chemical formula $C_{11}H_{21}N_5O_3$. Fura red ($C_{41}H_{44}N_4O_{20}S$) is the most challenging query for the model; it is ranked at 8,320 by All-Ensemble. Its entire description is based off of 1-benzofuran, but the substitutions are each larger than the original molecule and poorly defined.

5.5 Remaining Challenges

One further challenge is integrating external domain knowledge. Many current errors can be eliminated by applying this information, such as assuming "oxide" means the molecule should contain an oxygen. Although our association rule approach learns some of these, external knowledge can provide stronger rules. We observe that descriptions appear to be the limiting factor in this model, which is consistent with the similar performance of the GCN and MLP encoders. Comprehensive techniques for extracting information from external knowledge could lead to significant improvements, which we leave for future work.

6 Conclusions and Future Work

In this work, we present **Text2Mol**: a novel and challenging cross-modal information retrieval task to retrieve molecules using natural language descriptions. To tackle this problem, we apply contrastive representation learning to a BERT-based text encoder and both MLP and GCN-based molecule encoders. We show that these models are complementary and that an ensemble approach combines them very effectively. We also show that the ensemble approach is effective for combining identically trained neural networks (with different parameter initialization), and we consider attention-based association rules. Improved encoder architectures will likely yield improvements, and further investigation of how model architectural choices affect these rules and their interactions for reranking may be interesting as well. In the future, we plan to further improve results by integrating external knowledge as constraints. It should also be noted that this task is possible in the reverse direction, from molecules to descriptions. This has many possible applications, such as finding relevant descriptions for newly discovered molecules.

Acknowledgements

This research is based upon work supported by the Molecule Maker Lab Institute: An AI Research Institutes program supported by NSF under Award No. 2019897 and NSF No. 2034562. The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies, either expressed or implied, of the U.S. Government. The U.S. Government is authorized to reproduce and distribute reprints for governmental purposes notwithstanding any copyright annotation therein.

References

Derek T. Ahneman, Jesús G. Estrada, Shishi Lin, Spencer D. Dreher, and Abigail G. Doyle. 2018. Predicting reaction performance in c–n cross-coupling using machine learning. *Science*, 360(6385):186–190.

Jimmy Ba, Jamie Ryan Kiros, and Geoffrey E. Hinton. 2016. Layer normalization. *arXiv preprint arXiv:1607.06450*.

John M. Barnard. 1993. Substructure searching methods: Old and new. *Journal of chemical information and computer sciences*, 33(4):532–538.

Iz Beltagy, Kyle Lo, and Arman Cohan. 2019. SciBERT: A pretrained language model for scientific text. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3615–3620, Hong Kong, China. Association for Computational Linguistics.

Samarth Bharadwaj, Mayank Vatsa, and Richa Singh. 2014. Aiding face recognition with social context association rule based re-ranking. In *IEEE International Joint Conference on Biometrics*, pages 1–8. IEEE.

Paheli Bhattacharya, Pawan Goyal, and Sudeshna Sarkar. 2016. Using word embeddings for query translation for hindi to english cross language information retrieval. *Computación y Sistemas*, 20(3):435–447.

Adrià Cereto-Massagué, María J. Ojeda, Cristina Valls, Miquel Mulero, Santiago Garcia-Vallvé, and Gerard Pujadas. 2015. Molecular fingerprint similarity search in virtual screening. *Methods*, 71:58–63.

Yen-Chun Chen, Linjie Li, Licheng Yu, Ahmed El Kholy, Faisal Ahmed, Zhe Gan, Yu Cheng, and Jingjing Liu. 2019. Uniter: Learning universal image-text representations.

Seyone Chithrananda, Gabe Grand, and Bharath Ramsundar. 2020. Chemberta: Large-scale self-supervised pretraining for molecular property prediction. *arXiv preprint arXiv:2010.09885*.

Connor W. Coley, Regina Barzilay, Tommi S. Jaakkola, William H. Green, and Klavs F. Jensen. 2017. Prediction of organic reaction outcomes using machine learning. *ACS central science*, 3(5):434–443.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Harris Drucker, Corinna Cortes, Lawrence D. Jackel, Yann LeCun, and Vladimir Vapnik. 1994. Boosting and other ensemble methods. *Neural Computation*, 6(6):1289–1301.

Benedek Fabian, Thomas Edlich, Héléna Gaspar, Marwin Segler, Joshua Meyers, Marco Fiscato, and Mohamed Ahmed. 2020. Molecular representation learning with language models and domain-relevant auxiliary tasks. *arXiv preprint arXiv:2011.13230*.

Weiguo Fan, Michael Gordon, and Praveen Pathak. 2006. On linear mixture of expert approaches to information retrieval. *Decision Support Systems*, 42(2):975–987.

William Fedus, Barret Zoph, and Noam Shazeer. 2021. Switch transformers: Scaling to trillion parameter models with simple and efficient sparsity. *arXiv preprint arXiv:2101.03961*.

David Fooshee, Aaron Mood, Eugene Gutman, Mohammadamin Tavakoli, Gregor Urban, Frances Liu, Nancy Huynh, David Van Vranken, and Pierre Baldi. 2018. Deep learning for chemical reaction prediction. *Molecular Systems Design & Engineering*, 3(3):442–452.

Marta Glavatskikh, Jules Leguy, Gilles Hunault, Thomas Cauchy, and Benoit Da Mota. 2019. Dataset's chemical diversity limits the generalizability of machine learning predictions. *Journal of Cheminformatics*, 11(1):1–15.

Kunal Goyal, Utkarsh Gupta, Abir De, and Soumen Chakrabarti. 2020. Deep neural matching models for graph retrieval. In *Proceedings of the 43rd International ACM SIGIR conference on research and development in Information Retrieval, SIGIR 2020, Virtual Event, China, July 25-30, 2020*, pages 1701–1704. ACM.

Thomas R. Hagadone. 1992. Molecular substructure similarity searching: efficient retrieval in two-dimensional structure databases. *Journal of chemical information and computer sciences*, 32(5):515–521.

Jiawei Han and Jian Pei. 2000. Mining frequent patterns by pattern-growth: methodology and implications. *ACM SIGKDD explorations newsletter*, 2(2):14–20.

Janna Hastings, Gareth Owen, Adriano Dekker, Marcus Ennis, Namrata Kale, Venkatesh Muthukrishnan, Steve Turner, Neil Swainston, Pedro Mendes, and Christoph Steinbeck. 2016. Chebi in 2016: Improved services and an expanding collection of metabolites. *Nucleic acids research*, 44(D1):D1214–D1219.

Rezarta Islamaj Doğan, Sun Kim, Andrew Chatr-Aryamontri, Chih-Hsuan Wei, Donald C. Comeau, Rui Antunes, Sérgio Matos, Qingyu Chen, Aparna Elangovan, Nagesh C. Panyam, et al. 2019. Overview of the biocreative vi precision medicine track: mining protein interactions and mutations for precision medicine. *Database*, 2019.

Sabrina Jaeger, Simone Fulle, and Samo Turk. 2018. Mol2vec: unsupervised machine learning approach with chemical intuition. *Journal of chemical information and modeling*, 58(1):27–35.

Jeff Johnson, Matthijs Douze, and Hervé Jégou. 2017. Billion-scale similarity search with gpus. *arXiv preprint arXiv:1702.08734*.

Sunghwan Kim, Jie Chen, Tiejun Cheng, Asta Gindulyte, Jia He, Siqian He, Qingliang Li, Benjamin A. Shoemaker, Paul A. Thiessen, Bo Yu, et al. 2019. Pubchem 2019 update: improved access to chemical data. *Nucleic acids research*, 47(D1):D1102–D1109.

Sunghwan Kim, Paul A. Thiessen, Evan E. Bolton, Jie Chen, Gang Fu, Asta Gindulyte, Lianyi Han, Jane He, Siqian He, Benjamin A. Shoemaker, et al. 2016. Pubchem substance and compound databases. *Nucleic acids research*, 44(D1):D1202–D1213.

Diederik P. Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. In *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*.

Thomas N. Kipf and Max Welling. 2017. Semi-supervised classification with graph convolutional networks. In *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings*. OpenReview.net.

Martin Krallinger, Obdulia Rabal, Analia Lourenco, Julen Oyarzabal, and Alfonso Valencia. 2017. Information retrieval and text mining technologies for chemistry. *Chemical reviews*, 117(12):7673–7761.

Miroslav Kratochvíl, Jiří Vondrášek, and Jakub Galgonek. 2018. Sachem: a chemical cartridge for high-performance substructure search. *Journal of cheminformatics*, 10(1):1–11.

Miroslav Kratochvíl. 2019. Accelerating structure search in small-molecule databases. Master's thesis.

Tuan Lai, Heng Ji, ChengXiang Zhai, and Quan H. Tran. 2021. Joint biomedical entity and relation extraction with knowledge-enhanced collective inference. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 6248–6260, Online. Association for Computational Linguistics.

Greg Landrum. 2021. Rdkit: Open-source cheminformatics software.

Diya Li, Lifu Huang, Heng Ji, and Jiawei Han. 2019. Biomedical event extraction based on knowledge-driven tree-LSTM. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 1421–1430, Minneapolis, Minnesota. Association for Computational Linguistics.

Diya Li and Heng Ji. 2019. Syntax-aware multi-task graph convolutional networks for biomedical relation extraction. In *Proceedings of the Tenth International Workshop on Health Text Mining and Information Analysis (LOUHI 2019)*, pages 28–33, Hong Kong. Association for Computational Linguistics.

Hongzhi Li, Joseph G. Ellis, Heng Ji, and Shih-Fu Chang. 2016. Event specific multimodal pattern mining for knowledge base construction. In *Proceedings of the 2016 ACM Conference on Multimedia Conference, MM 2016, Amsterdam, The Netherlands, October 15-19, 2016*, pages 821–830.

Manling Li, Alireza Zareian, Qi Zeng, Spencer Whitehead, Di Lu, Heng Ji, and Shih-Fu Chang. 2020. Cross-media structured common space for multimedia event extraction. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 2557–2568, Online. Association for Computational Linguistics.

Daniel M. Lowe, Peter T. Corbett, Peter Murray-Rust, and Robert C. Glen. 2011. Chemical name to structure: Opsin, an open source solution. *Journal of Chemical Information and Modeling*, 51(3):739–753. PMID: 21384929.

Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. 2019. Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. In *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada*, pages 13–23.

Emily K. Mallory, Ambika Acharya, Stefano E. Rensi, Peter J. Turnbaugh, Roselie A. Bright, and Russ B. Altman. Chemical reaction vector embeddings: towards predicting drug metabolism in the human gut microbiome. In *Biocomputing 2018*, pages 56–67.

Saeed Masoudnia and Reza Ebrahimpour. 2014. Mixture of experts: a literature survey. *Artificial Intelligence Review*, 42(2):275–293.

Christian Meilicke, Melisachew W. Chekol, Manuel Fink, and Heiner Stuckenschmidt. 2020. Reinforced anytime bottom up rule learning for knowledge graph completion. *arXiv preprint arXiv:2004.04412*.

Christian Meilicke, Melisachew W. Chekol, Daniel Ruffinelli, and Heiner Stuckenschmidt. 2019. Anytime bottom-up rule learning for knowledge graph completion. In *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence, IJCAI 2019, Macao, China, August 10-16, 2019*, pages 3137–3143. ijcai.org.

Tomás Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013a. Efficient estimation of word representations in vector space. In *1st International Conference on Learning Representations, ICLR 2013, Scottsdale, Arizona, USA, May 2-4, 2013, Workshop Track Proceedings*.

Tomás Mikolov, Ilya Sutskever, Kai Chen, Gregory S. Corrado, and Jeffrey Dean. 2013b. Distributed representations of words and phrases and their compositionality. In *Advances in Neural Information Processing Systems 26: 27th Annual Conference on Neural Information Processing Systems 2013. Proceedings of a meeting held December 5-8, 2013, Lake Tahoe, Nevada, United States*, pages 3111–3119.

Aditya Mogadala, Marimuthu Kalimuthu, and Dietrich Klakow. 2020. Trends in integration of vision and language research: A survey of tasks, datasets, and methods. *arXiv preprint arXiv:1907.09358*.

Simon Ott, Christian Meilicke, and Matthias Samwald. 2021. SAFRAN: An interpretable, rule-based link prediction method outperforming embedding models. In *3rd Conference on Automated Knowledge Base Construction*.

Robi Polikar. 2012. Ensemble learning. In *Ensemble machine learning*, pages 1–34. Springer.

Jingwei Qu, Penghui Sun, Xin Li, Bei Wang, Xiaoqing Lu, Zhi Tang, and Chengcui Zhang. 2019. A retrieval system of medicine molecules based on graph similarity. *IEEE MultiMedia*, 26(4):17–27.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. 2021. Learning transferable visual models from natural language supervision. In *Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event*, volume 139 of

Proceedings of Machine Learning Research, pages 8748–8763. PMLR.

Stefano Rensi and Russ B. Altman. 2017. Flexible analog search with kernel pca embedded molecule vectors. *Computational and structural biotechnology journal*, 15:320–327.

Tim Rocktäschel, Torsten Huber, Michael Weidlich, and Ulf Leser. 2013. WBI-NER: The impact of domain-specific features on the performance of identifying and classifying mentions of drugs. In *Second Joint Conference on Lexical and Computational Semantics (*SEM), Volume 2: Proceedings of the Seventh International Workshop on Semantic Evaluation (SemEval 2013)*, pages 356–363, Atlanta, Georgia, USA. Association for Computational Linguistics.

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

Frederik Sandfort, Felix Strieth-Kalthoff, Marius Kühnemund, Christian Beecks, and Frank Glorius. 2020. A structure-based platform for predicting chemical reactivity. *Chem*, 6(6):1379–1390.

Kihyuk Sohn. 2016. Improved deep metric learning with multi-class n-pair loss objective. In *Advances in Neural Information Processing Systems 29: Annual Conference on Neural Information Processing Systems 2016, December 5-10, 2016, Barcelona, Spain*, pages 1849–1857.

Weijie Su, Xizhou Zhu, Yue Cao, Bin Li, Lewei Lu, Furu Wei, and Jifeng Dai. 2020. VL-BERT: pre-training of generic visual-linguistic representations. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, pages 5998–6008.

Ivan Vulic and Marie-Francine Moens. 2015. Monolingual and cross-lingual information retrieval models based on (bilingual) word embeddings. In *Proceedings of the 38th International ACM SIGIR Conference on Research and Development in Information Retrieval, Santiago, Chile, August 9-13, 2015*, pages 363–372. ACM.

David Weininger. 1988. Smiles, a chemical language and information system. 1. introduction to methodology and encoding rules. *Journal of chemical information and computer sciences*, 28(1):31–36.

David Weininger, Arthur Weininger, and Joseph L. Weininger. 1989. Smiles. 2. algorithm for generation of unique smiles notation. *Journal of chemical information and computer sciences*, 29(2):97–101.

Zhenqin Wu, Bharath Ramsundar, Evan N. Feinberg, Joseph Gomes, Caleb Geniesse, Aneesh S. Pappu, Karl Leswing, and Vijay Pande. 2018. Moleculenet: a benchmark for molecular machine learning. *Chemical science*, 9(2):513–530.

Xifeng Yan, Philip S. Yu, and Jiawei Han. 2005. Substructure similarity search in graph databases. In *Proceedings of the 2005 ACM SIGMOD international conference on Management of data*, pages 766–777.

Liang Zhang and Xiaobing Zhao. 2020. An overview of cross-language information retrieval. In *International Conference on Artificial Intelligence and Security*, pages 26–37. Springer.

Zixuan Zhang, Nikolaus Parulian, Heng Ji, Ahmed Elsayed, Skatje Myers, and Martha Palmer. 2021. Fine-grained information extraction from biomedical literature based on knowledge-enriched Abstract Meaning Representation. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 6261–6270, Online. Association for Computational Linguistics.

Jin Guang Zheng, Daniel Howsmon, Boliang Zhang, Juergen Hahn, Deborah McGuinness, James Hendler, and Heng Ji. 2014. Entity linking for biomedical literature. In *BMC Medical Informatics and Decision Making*.

Yingyao Zhou, Bin Zhou, Shumei Jiang, and Frederick J. King. 2010. Chemical- text hybrid search engines. *Journal of chemical information and modeling*, 50(1):47–54.

A Supporting Figures

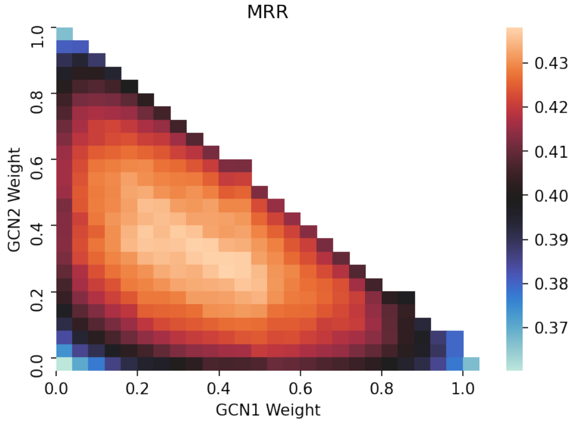


Figure 7: This figure shows the ensemble validation MRR from different weighted averages of the three GCN models. GCN3_weight = 1 − GCN1_weight − GCN2_weight. The MRR is clearly lower in the corners, where only rankings from one model are used because the others have zero weight. This figure illustrates that using uniform weights is an effective heuristic.

B Reproducibility

The MLP and GCN models were each run three times. The GCN and MLP use 600 hidden units. The mol2vec input and the model outputs are 300-dimensional. GCN uses the substructure representations with the largest radius. MLP contains 110,871,865 parameters. GCN contains 111,953,665 parameters. The cross-modal attention model contains 128,978,441 parameters and attends the first 512 molecule substructures. It achieves about 97% classification accuracy for the matching task from the negative samples. The number of one-to-one association rules with confidence greater than 0.1 and support greater than 2 is 1,835. The MLP and GCN take approximately 7 hours on a NVIDIA V100 and the cross-modal attention model takes approximately 9 hours. We find that early stopping is not useful and that layer normalization increases training speed. The value of α for reranking was selected by grid search for high validation MRR. For the metrics, given a list of rankings R,

$$MeanRank = \frac{1}{n}\sum_{i=1}^{n} R_i$$

$$MRR = \frac{1}{n}\sum_{i=1}^{n} \frac{1}{R_i}$$

	Validation			
Model	Mean Rank	MRR	Hits@1	Hits@10
MLP1	43.66	0.374	22.5%	68.8%
MLP2	47.42	0.360	22.1%	68.9%
MLP3	41.15	0.376	21.2%	68.2%
GCN1	41.78	0.367	22.2%	68.4%
GCN2	41.23	0.367	22.1%	68.9%
GCN3	42.19	0.360	21.2%	68.2%
MLP-Ensemble	30.60	0.442	28.7%	76.5%
GCN-Ensemble	28.89	0.435	27.7%	76.6%
All-Ensemble	**24.95**	**0.479**	**31.7%**	**80.2%**

Table 4: Reproducibility results for the validation set.

$$Hits@m = \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}_{R_i \leq m}$$