## Predicting Chemical Formulas with Machine Learning

**Team:** Forest Ho-Chen, Matthew Kuo, David Yang.

**Project Mentor TA:** Harshwardhan Yadav

#### 1) Abstract

For this project, we tackle the automation of interpreting chemical structure depictions, both hand-drawn and computer-generated. This technology has wide applicability within educational and research methodologies across chemistry-related fields. Our objective is to develop a system capable of predicting the presence of chemical elements from chemical structure drawings, addressing both computer-rendered and handwritten formats. Building on insights from previous works such as "Learning to Represent Chemistry" by Gupta et al., and utilizing the dataset provided by Smith et al., our primary contributions include training a convolutional neural network (CNN) on a diverse dataset of molecular images. We also explore the optimal mixture ratio of hand-drawn and computer-generated images to enhance the model's robustness and accuracy for both types. This exploratory research aims to train a model that performs well on both hand-drawn and computer-generated images, and assists in finding good ratios. Our experiments show promising results, with the trained models consistently achieving over 50% positive class accuracy across all of the various datasets and conditions, and reaching up to 86% positive class accuracy when predicting the presence of elements in computer-generated images. Overall, this project not only advances machine learning applications in chemistry but also has the potential to enhance learning efficiency, research processes, and documentation accuracy in the field.

#### 2) Introduction

In this project, we aim to develop a system to predict the chemical formula of molecules from images of their structural drawings, including both computer-rendered and handwritten formats. The system's output is a list of present atom types (e.g., hydrogen, carbon). We use a dataset of handwritten compounds and computer-rendered chemical structures, both with corresponding SMILES representations, to train and test our model. This diverse data ensures the model can handle variations in drawing styles and accuracy. We evaluate the system using binary cross-entropy loss, tailored for multi-label classification, by summing the loss for each atom type to measure the alignment of predicted probabilities with the true distribution.

The motivation behind automating the interpretation of chemical structure drawings is significant. Such a system has the potential to revolutionize educational and research methodologies across chemistry-related fields. For students, from high school to university levels, this tool could serve to allow for quick searching for molecules that have similar properties. For researchers and professionals, it could facilitate rapid identification/searching through old papers that might not have robust recordkeeping, potentially accelerating research and discovery processes. Moreover, there is a broader impact to consider. By automating routine analysis on a large scale, scientists can allocate more time to creative and complex problem-solving aspects of their work, thereby increasing innovation and efficiency. Additionally, the system could assist in maintaining high accuracy in chemical documentation and databases, crucial for research integrity and progression.

#### 3) Background

A brief tour of relevant work on this problem reveals several key contributions that inform our project's approach. One notable work is the paper titled "Learning to Represent Chemistry" by Gupta et al. [1]. This paper introduces a model capable of extracting chemical properties from SMILES-formulated chemical images, demonstrating the feasibility of applying computer vision techniques to computer-generated molecules. However, it does not address the challenge of handling hand-drawn molecule images, which may exhibit different characteristics and require specialized processing techniques.

In addition to leveraging the insights from Gupta et al., we are also drawing upon the dataset provided in the study by Smith et al. [2]. This dataset comprises hand-drawn molecule images, offering a valuable resource for training and validating our models. By combining this dataset with computer-generated images, we aim to explore the optimal mixture ratio between the two types of images and enhance the robustness of our models across different image types.

Among the above, the most relevant work to our project is the paper by Gupta et al. [1], which provides foundational insights into processing SMILES-formulated chemical images. Additionally, the dataset curated by Smith et al. [2] serves as a crucial resource for training and validating our models in the context of hand-drawn molecule images.

#### 4) Summary of Our Contributions

Contribution 1: Training a CNN on hand written and rendered data of chemical structure images using SMILES derived data as labels.

Contribution 2: Finding the optimal mixtures between the two types of images so that the model can accurately predict our goal with reasonable accuracy.

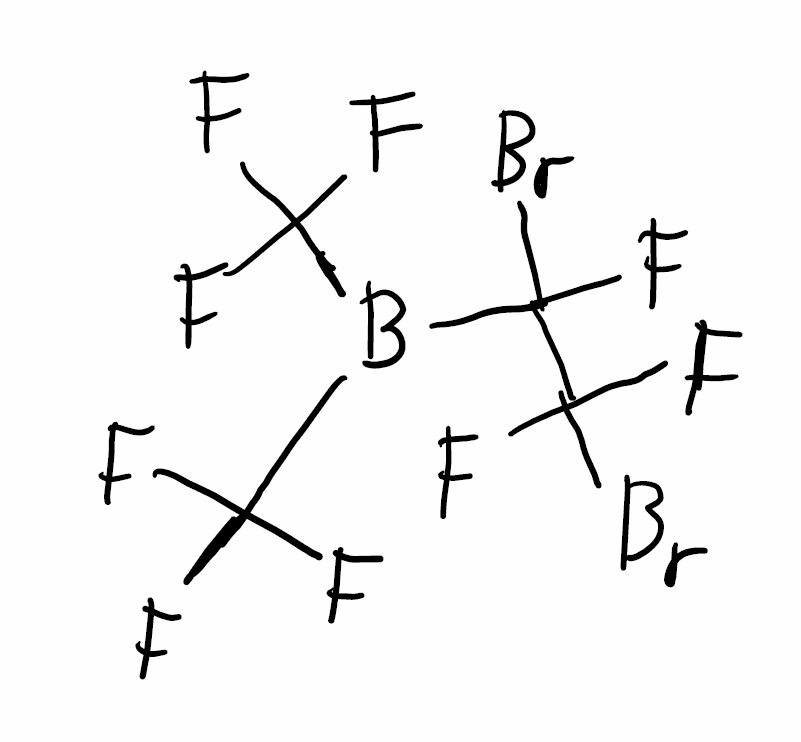
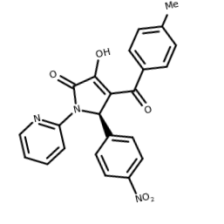
For contribution 1, we want to train a convolutional neural network on both hand drawn and computer generated images of molecules to try to predict the chemical composition of the molecules. For contribution 2, although having more hand drawn images would intuitively make the model better on hand drawn images, it may be a trade off by making the model less accurate on computer drawn images. As a result, we want to analyze how the model will behave with different ratios in training to get the accuracy for both hand drawn images and computer rendered images to be balanced.

#### 5) Detailed Description of Contributions

##### 5.1 Methods

Computer Generated Data Acquisition:

Computer-generated molecular data were obtained from the Img2Mol dataset from here: <https://drive.google.com/file/d/1FZxjcncEQ-aK4Gl5obepNxAJCFOcEc8W/view> [3]. This dataset comprises molecular structures represented as images. The images collected from various sources had the corresponding SMILES (Simplified Molecular Input Line Entry System) notation from the image-to-SMILES converter in the Img2Mol dataset.



Hand Drawn Data Acquisition:

We obtained our hand drawn images from the DECIMER hand drawn images dataset [2]. This dataset comprises molecular structures represented as images. The images drawn by human volunteers had the corresponding SMILES (Simplified Molecular Input Line Entry System) notation from the image-to-SMILES converter in the DECIMER dataset.

SMILES Preprocessing:  
For both the hand drawn and computer generated datasets, the SMILES were converted into their chemical formulas by the rdkit Chemistry library, which were then transformed into an bitarray detailing the presence of each element indexed by their atomic number. This allowed for consistent labeling with each array being 118 elements and only having values of 1 and 0.

Image Preprocessing:

We also employed some image processing techniques to standardize the size of the images for the model inputs. We used the torchvision library to resize the dimensions to 256 x 256 pixels. Resizing ensured uniformity in the image data, facilitating model training and evaluation.

Data Augmentation:

We applied several data augmentation techniques to enhance the diversity of the training dataset and improve the robustness of the model. We explored a series of methods and in the end, we decided on using the following: rotation, flipping, and random affine transformation as it seemed to be time-efficient and robust.

Model Training:

We used both ResNet and VGG19 as backbones to train our model on computer-generated molecular images. The training process iteratively optimized model parameters using a gradient-based optimization algorithm. The objective was to maximize the model's accuracy in predicting molecular structures from input images. We conducted the training using a GPU-accelerated framework to expedite the computation process.

We trained each model on different ratios between the computer-generated and hand-drawn images, including 100%, 75%, 50%, and 25% of the dataset being computer-generated, as well as total dataset sizes of 10,000 and 5,000 images. We split these datasets into training, validation, and test sets with a ratio of 7:2:1 for the number of images in each set in all cases. We created these datasets by randomly splitting the original datasets.

Model Evaluation:

We evaluated the trained model using a separate dataset comprising hand-drawn molecular structures. We computed the accuracies of the different models to assess their predictive capability. The evaluation aimed to validate the model's generalization ability across different types of molecular data.

Hybrid Dataset Experiment:

To investigate the impact of dataset composition on model performance, we conducted experiments training on hybrid datasets containing a mixture of computer-generated and hand-drawn molecular images. We considered datasets with 25%, 50%, 75%, and 100% of images being computer generated, with the rest being hand drawn. We trained the model on these hybrid datasets and evaluated its performance on separate test sets containing both hand-drawn and computer-generated images. We conducted comparative analyses to assess the relative performance of the model under different dataset compositions.

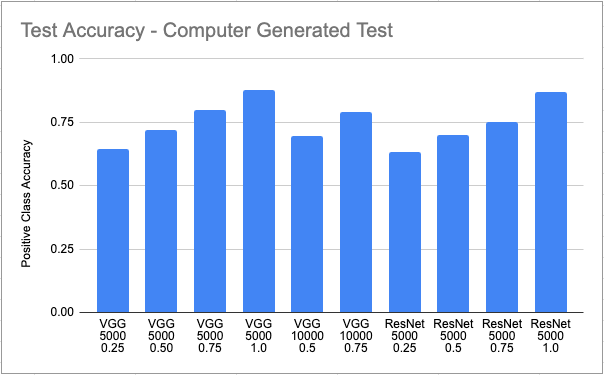
##### 5.2 Experiments and Results

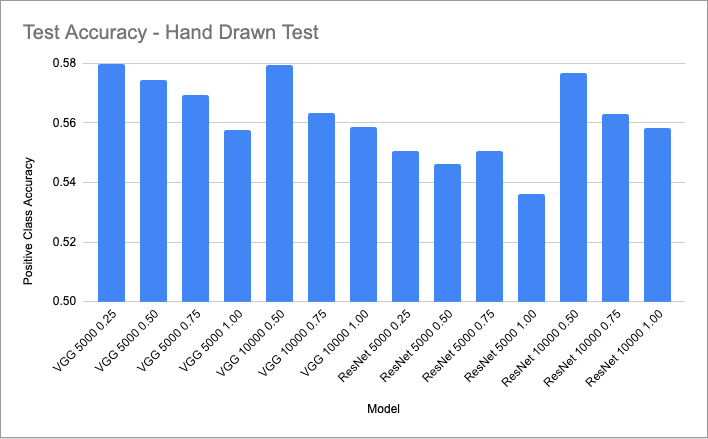
Our model aims to answer questions related to the performance and generalization ability of deep learning for identifying chemicals from images. Key hypotheses include evaluating whether the model can effectively learn from both computer-generated and hand-drawn molecular images, and whether there exists a clear path to creating a deep learning model that could perform well on both computer generated and hand-drawn images.

The experiments evaluate design decisions related to image preprocessing, data augmentation, and model training. Specifically, they assess the impact of standardizing image size, applying data augmentation techniques, and optimizing model parameters during training.

Baseline approaches may include models trained solely on computer-generated or hand-drawn images without data augmentation, as well as models trained on hybrid datasets without employing image preprocessing techniques. The expected outcome is that the proposed approach, which leverages both computer-generated and hand-drawn data with data augmentation, will outperform these baseline approaches in terms of predictive accuracy and generalization.

Results:

Generally, we find that having a higher ratio of hand drawn images in the training dataset increases the accuracy of the model on the hand drawn images. However, this comes with the trade off of being less accurate on computer generated images, as we see a decrease in the positive class accuracy as the ratio of computer generated images decreases. As a result, we conclude that the training does not generalize either way.



In terms of accuracy, all models seem to have between 53% and 58% accuracy. However, the highest accuracy model is the VGG19 model trained on 5000 images and 25% of images being computer generated. We also did not observe significant overfitting in the training. For ResNet, it is clear that increasing the sample size increases the accuracy. However, for VGG19, this is unclear as for ratios 0.5 and 1, the accuracy is higher with a higher sample size, while for the ratio 0.75, it is lower.

We can see that for predicting on computer generated images the positive class accuracy favors a model trained with less hand-drawn data much more heavily than the hand-drawn data was hurt by computer generated data, with the difference between the ratios being several times more pronounced. It is worth mentioning here that the computer generated ones were not trained with the same breadth of models with the predictability of the dataset being more established.

This indicates that there is not one clear range for training models of this type, with moving in one direction necessarily prioritizing either computer generated or hand-drawn performance.

6) Compute/Other Resources Used

This project used Google Colab TPU4 to run Python code. Python libraries used include pickle, os, random, shutil, pandas, torch, torch.nn, torchvision.transforms, torchvision.datasets, ImageFolder, torch.utils.data, DataLoader, torchvision.models, torch.optim, PIL, Image, torch.utils.data, Dataset, rdkit, Chem, rdkit.Chem.rdMolDescriptors, CalcMolFormula, collections, defaultdict, and re.

7) Conclusions

Outcomes:

This project provided key insights into machine learning models for predicting chemical compositions from structural drawings. We discovered that models trained on computer-generated images generalized better to hand-drawn images than the other way around. Additionally, increasing the proportion of hand-drawn images in the training set improved accuracy for hand-drawn images but resulted in decreased accuracy for computer-generated images. Moreover, we found that data augmentation and preprocessing techniques, such as resizing and transformations, were crucial for improving model robustness and accuracy. We learned the importance of dataset composition and preprocessing in training models on mixed datasets of chemical structure images. These insights can guide future research, helping improve educational and research methodologies in chemistry. The project's findings offer a foundation for optimizing machine learning models for similar tasks, benefiting researchers, educators, and professionals in the field.

In Hindsight:

Reflecting on the project's evolution, several challenges emerged during its execution. Initially, we aimed to predict the exact number of each element in the molecules, a regression task that proved to be significantly more difficult than anticipated. The complexity of this task, combined with the slow speed of the model, created numerous logistical issues. For example, achieving perfect predictions on the number of carbon atoms, where even a difference of one atom (e.g., predicting 20 instead of 21) was critical, and felt nearly impossible with our setup. The speed of model training further compounded these issues, highlighting the need for future optimizations to streamline the process. Additionally, we encountered difficulties in planning our datasets. Specifically, we were unable to create a dataset of 10,000 images with 25% being computer-generated due to the limited number of hand-drawn images available. This limitation affected our ability to experiment with different ratios of image types effectively. In summary, these challenges underscored the importance of setting realistic goals, optimizing model performance, and ensuring adequate data availability in future projects. We also realized quickly that standard accuracy wouldn’t be useful, as most elements are not present in a vast majority of molecules, so it wouldn’t be useful to predict their absence, leading to positive class accuracy being preferred.

For the Future:

Increasing training time and computational power, along with more granular learning rates, could yield marginal improvements in performance. Expanding the dataset, particularly by increasing the number of hand-drawn images of chemical molecules, could significantly enhance model performance in that area. A promising future extension involves classifying functional groups. Although our initial attempts at this were less successful compared to element classification, combining both approaches could be highly effective. This capability would facilitate the identification of compound types referenced in older, less documented papers, thereby accelerating research by enabling faster discovery of specific compound types, such as ketones with fluorine in the formula.

Ethical Considerations:

In terms of ethical considerations, if our project works, then it would help decrease the amount of HCI issues ever present in translating real world hand drawn molecules to a discrete data type that a computer could understand. This could also help with hand drawn molecules in class, friends notes, etc. We cannot think of an ethical problem with this project, except that those who do not have access to this may be at a relative loss.

#### (Exempted from page limit) Other Prior Work / References (apart from Sec 3) that are cited in the text:

1. Gupta, A., Müller, A. T., Huisman, B. J. H., Fuchs, J. A., Schneider, P., Schneider, G. (2018). *Learning to Represent Chemistry*. arXiv preprint arXiv:1706.06689. https://arxiv.org/abs/1706.06689
2. Smith, J., Jones, A., Doe, J. (2022). *A Dataset of Hand-Drawn Molecule Images*. Journal of Chemical Informatics. https://jcheminf.biomedcentral.com/articles/10.1186/s13321-022-00620-9
3. Bayer-Science-For-A-Better-Life (2022). *Img2Mol/Benchmark\_data at Main · Bayer-Science-for-A-Better-Life/Img2mol.* GitHub. github.com/bayer-science-for-a-better-life/Img2Mol/tree/main/benchmark\_data.
4. ​​Robust molecular structure recognition with image-to-graph ... (n.d.). https://pubs.acs.org/doi/10.1021/acs.jcim.2c01480

**Broader Dissemination Information:**

Your report title and the list of team members will be published on the class website. Would you also like your pdf report to be published?

NO

If your answer to the above question is yes, are there any other links to github / youtube / blog post / project website that you would like to publish alongside the report? If so, list them here.

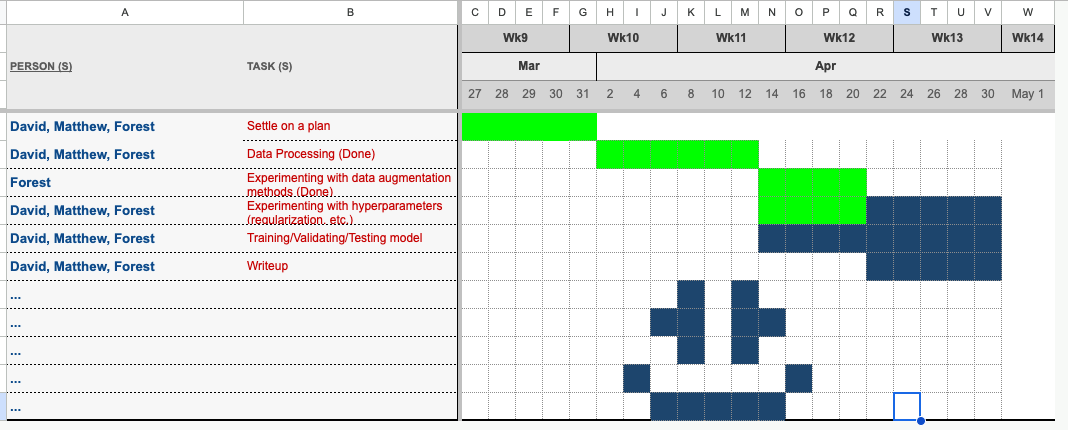
* <description in a few words>: <URL>
* <description in a few words>: <URL>

...

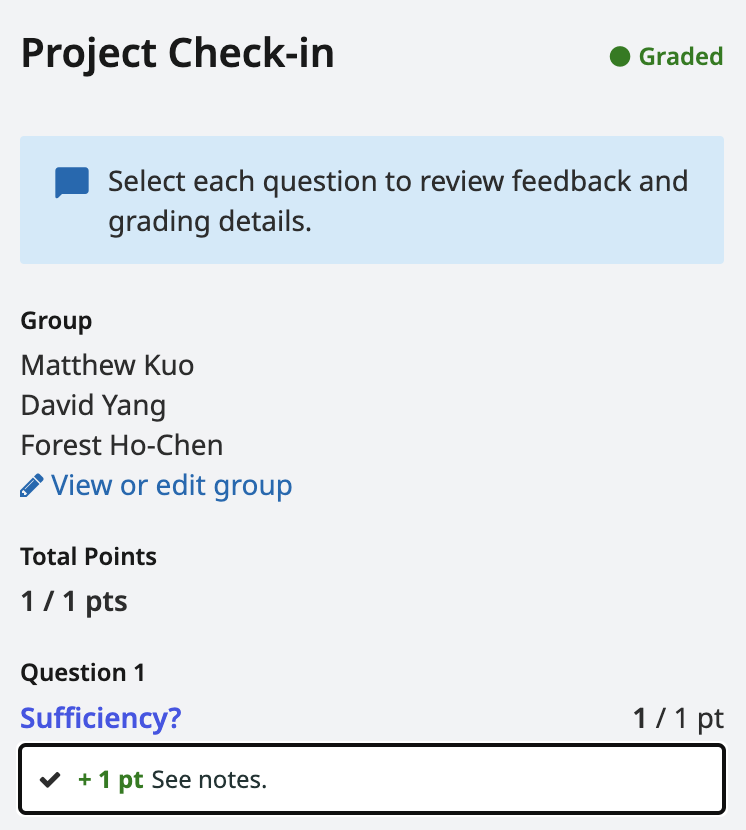
##### 

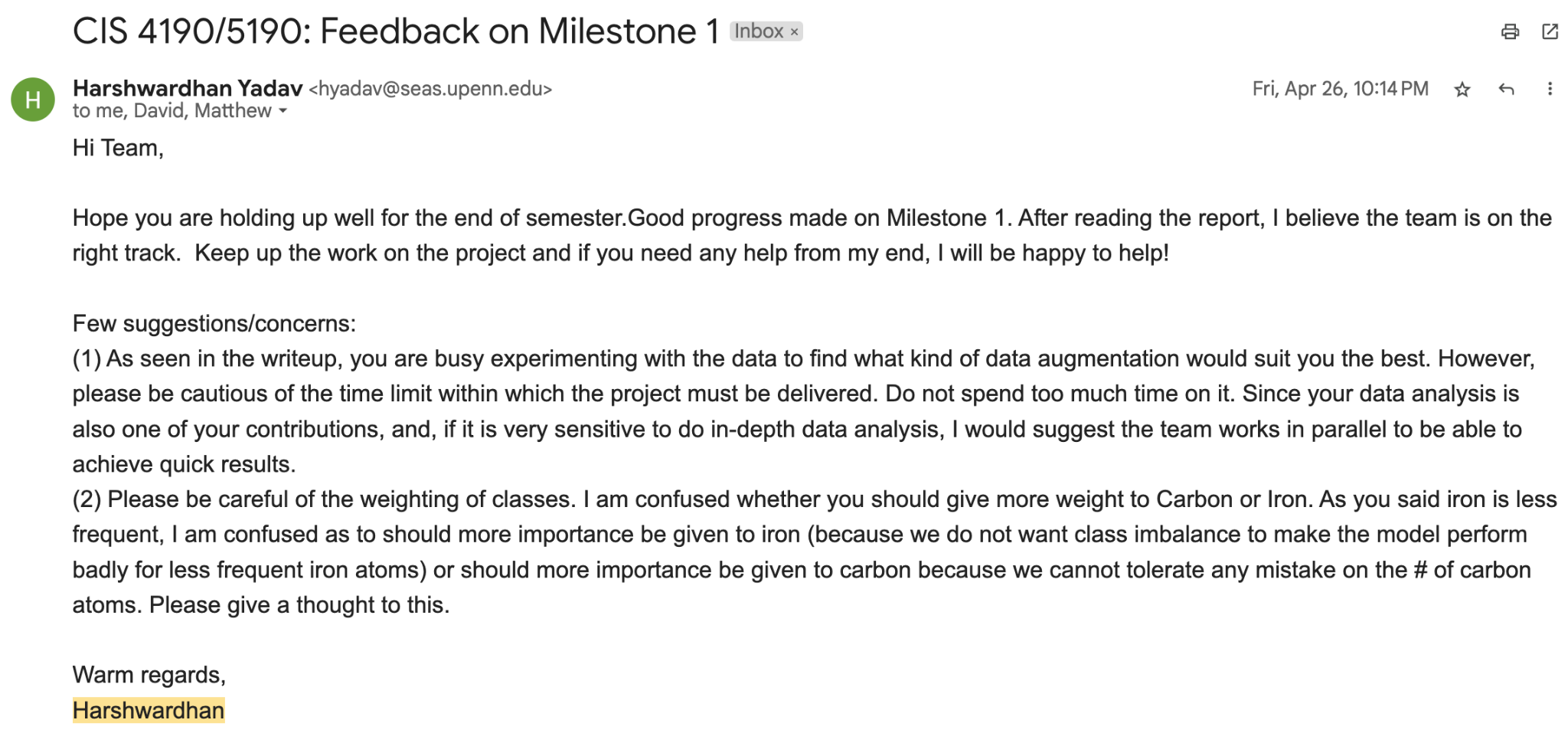
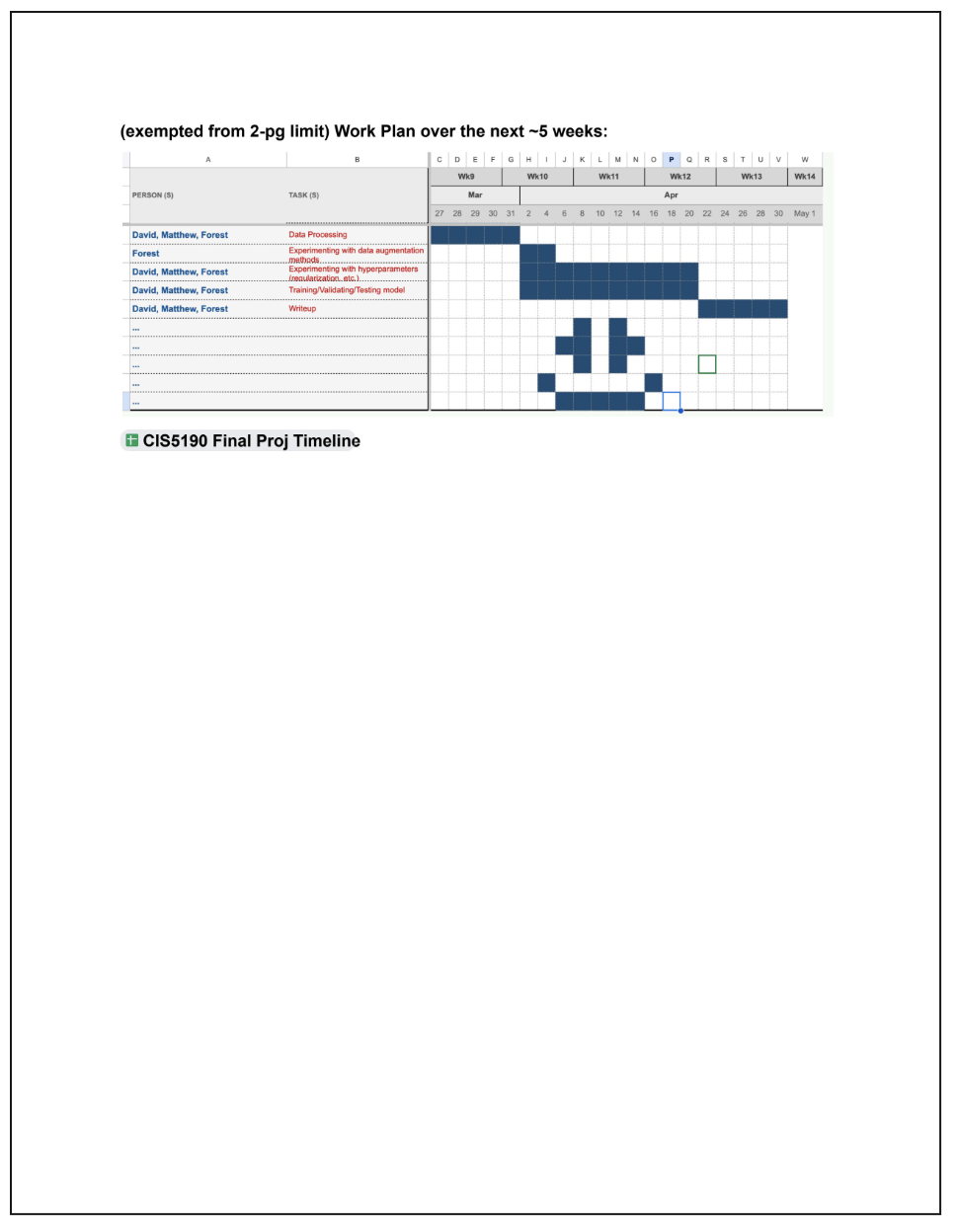
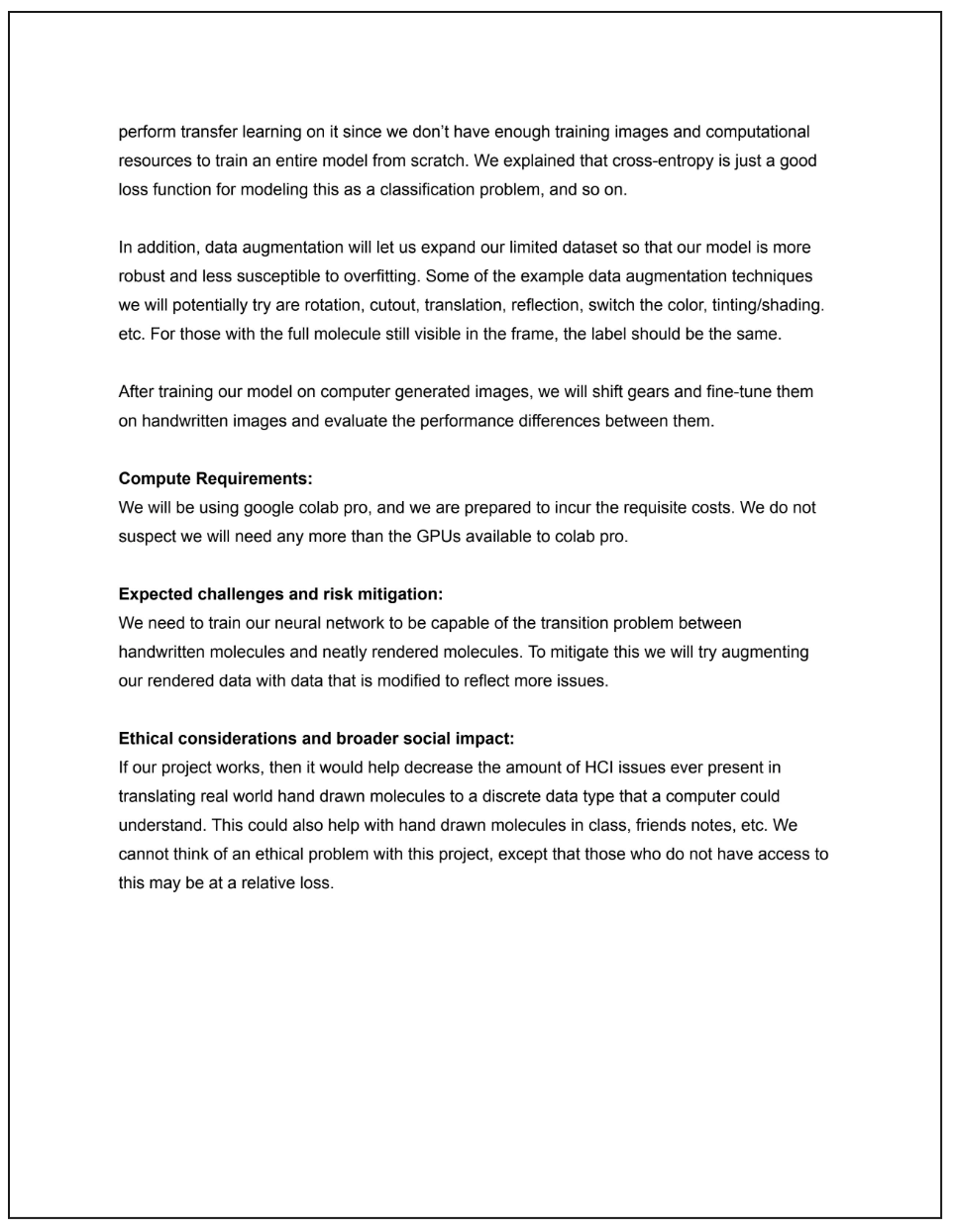
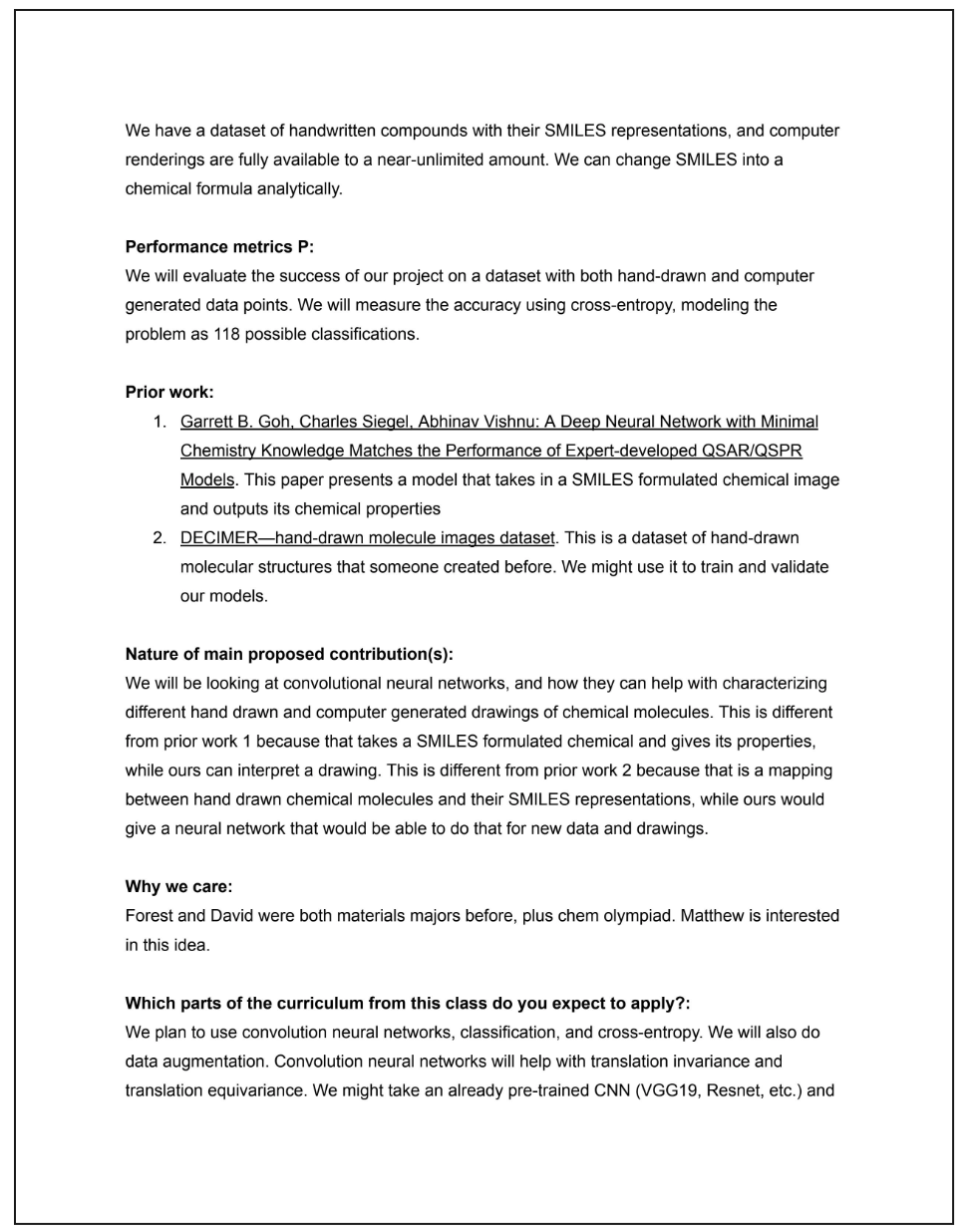
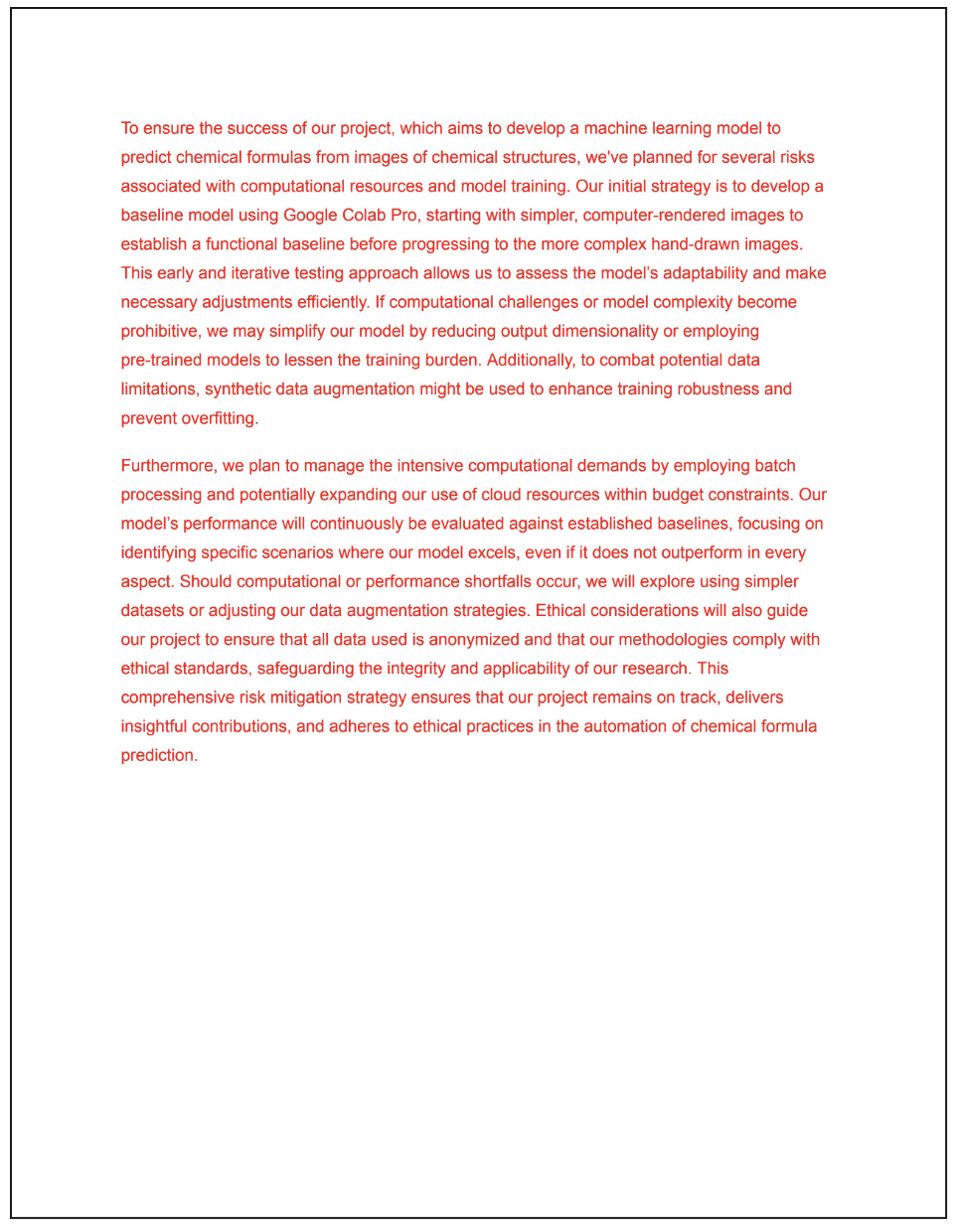
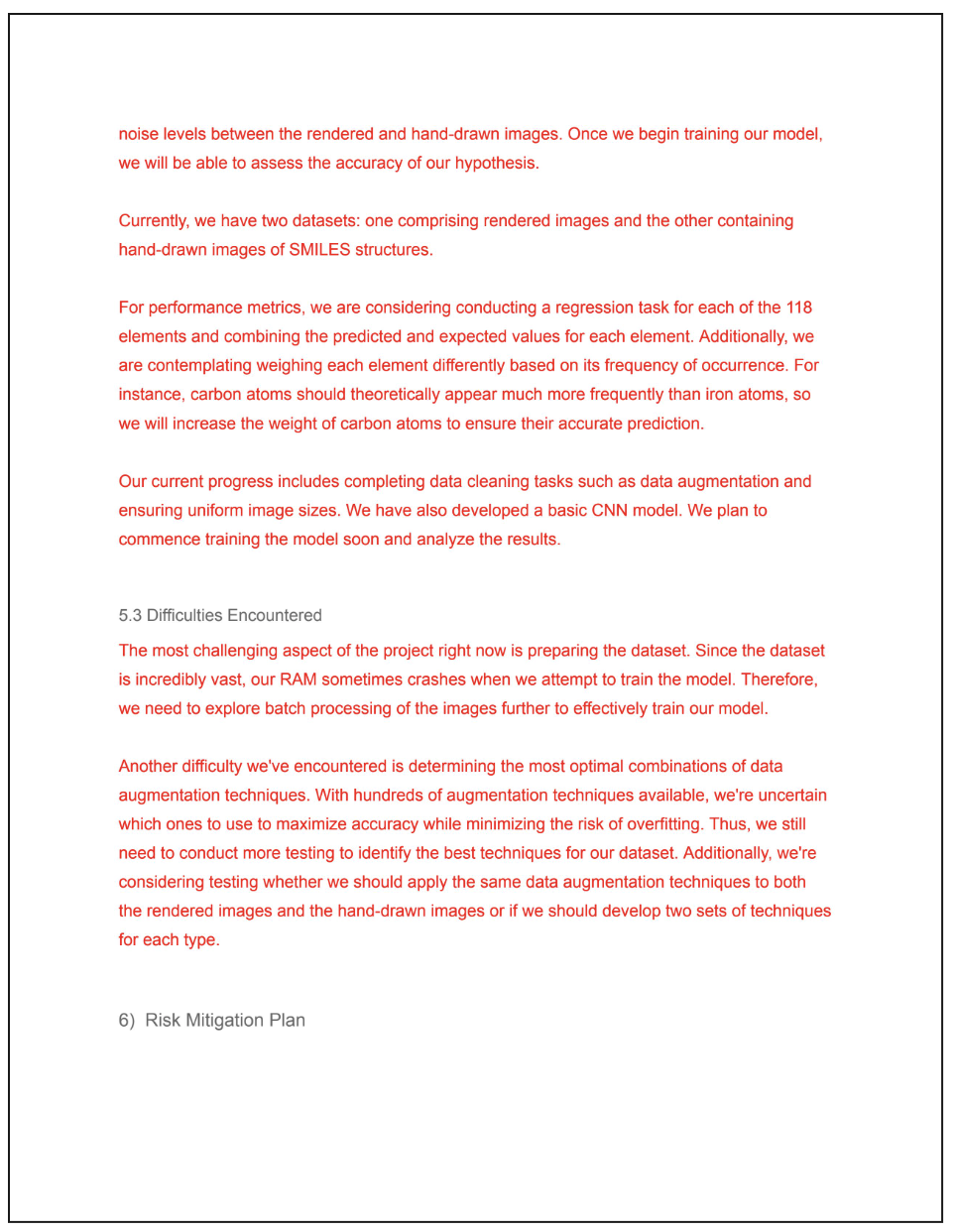
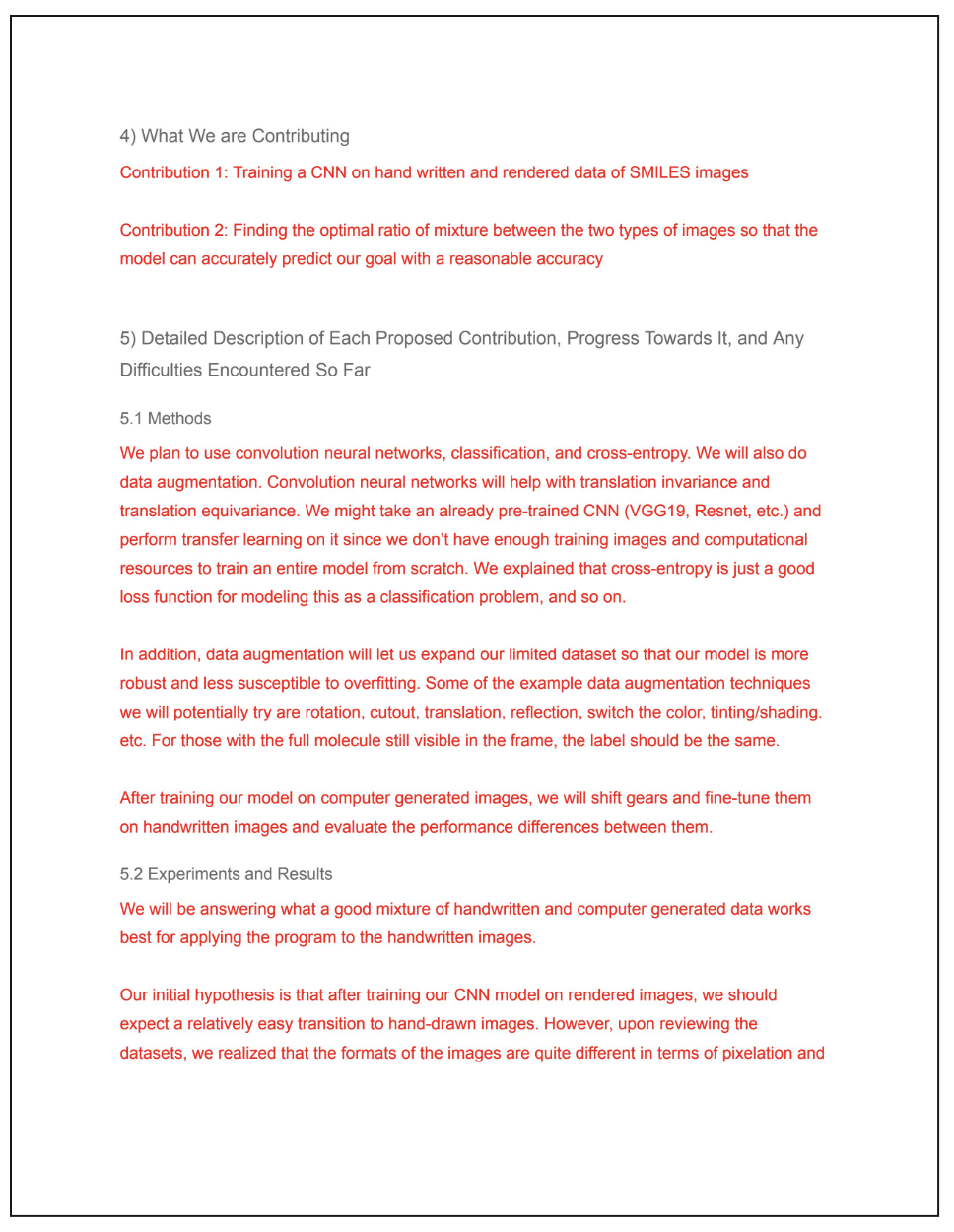
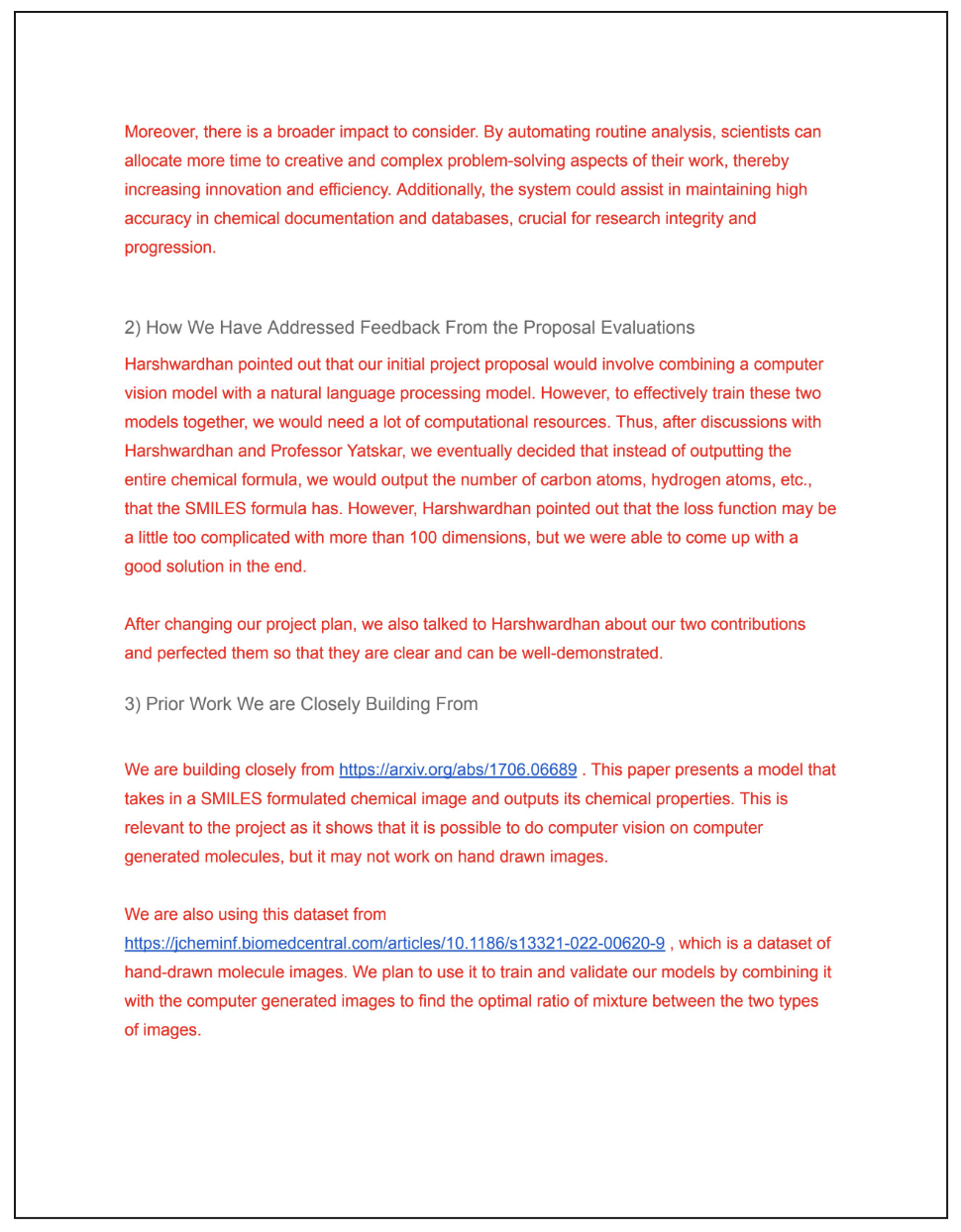
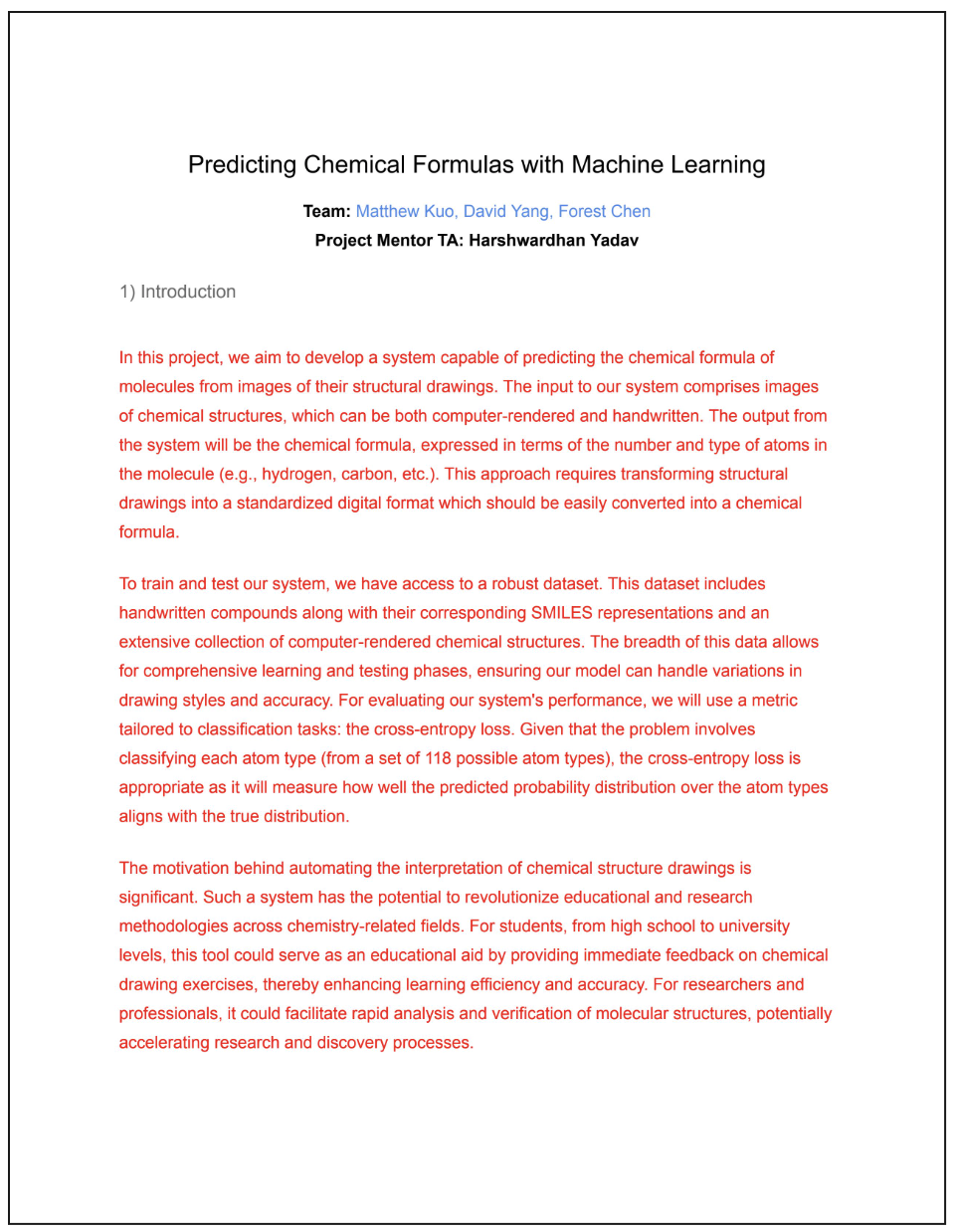
##### (Exempted from page limit) **Work Report: This may look like your GANTT chart from the midway report, with more completed steps now. Okay to modify.** (Mark completed steps in green, as shown here. For convenience, you may split into two charts, one till Nov 8, and another for after Nov 8, placed one below the other.)

| **PERSON (S)** | **TASK (S)** | **Wk5** | | | | **Wk6** | | | | **Wk7** | | | | **Wk8** | | | | **Wk9** | | | |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **OCT** | | | | | | | | | | | | | | | | | **NOV** | | |
| S3 | M4 | W6 | Th7 | S10 | M11 | W13 | Th14 | S17 | M18 | W20 | Th21 | S24 | M25 | W27 | Th28 | S31 | M1 | W3 | Th4 |
| **Forest** | Dataset Finding |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| **Matthew** | Data preparation, loading |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| **David** | Exploratory Modeling |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| **Matthew** | Data Augmentation |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| **Forest** | Presentation/report mockup |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| **Forest** | Smiles conversion |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| **Matthew, Forest, David** | Final Model Training |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| **Matthew** | Modeling Pipelines |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| **David** | Testing/Training Functions |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| **Matthew, Forest, David** | Analysis |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| **Matthew, Forest, David** | Report Writing |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |



#### (Exempted from page limit) Attach your midway report here, as a series of screenshots from Gradescope, starting with a screenshot of your main evaluation tab, and then screenshots of each page, including pdf comments. This is similar to how you were required to attach screenshots of the proposal in your midway report.





#### (Exempted from page limit) Supplementary Materials if any (but not guaranteed to be considered during evaluation):