Reverse Discovery of Sustainable Aviation Fuels Based on TenGAN

Karl I. Neivelt

Abstract

The aviation industry is responsible for 3% of global greenhouse‑gas emissions. Sustainable aviation fuels (SAFs) can close part of the gap, yet their molecular design space spans possibilities, making brute‑force experimental discovery pointless. Here, we adapt the Transformer-Encoder Generative Adversarial Network (*TenGAN*) architecture to generate candidate SAF molecules, followed by rapid property screening using group‑contribution estimators and genetic optimisation. Predictive uncertainties are quantified with conformal prediction. Finally, first‑principles density functional theory (DFT) validates combustion‑relevant descriptors for the top ten molecules. The workflow illustrates how generative adversarial networks aid in property‑oriented fuel discovery.

Some journals require a graphical entry for the Table of Contents. This should be laid out “print ready” so that the sizing of the text is correct.

Inside the tocentry environment, the font used is Helvetica 8 pt, as required by *Journal of the American Chemical Society*.

The surrounding frame is 9 cm by 3.5 cm, which is the maximum permitted for *Journal of the American Chemical Society* graphical table of content entries. The box will not resize if the content is too big: instead it will overflow the edge of the box.

This box and the associated title will always be printed on a separate page at the end of the document.

# Introduction

3-4 paragraphs with 20-30 references. Take lots from literature review  
P1: SAFs are important, but finding suitable molecules with wanted properties is hard. Experiments take a long time and lots of resouces.  
P2: Recent developments in machine learning methods and computer hardware allow for inverse design of SAFs. Although various generative AI model arhcitectures, such as VAEs, GPTs, have foud usage, generative adversarial networks have found the widest traction . These authors have developed models to generate novel drug molecules. These authors have developed models to generate novel fuel molecules. This is why they are not ideal.  
P3: Here, we present modified version of TenGAN , that has been adapted to the wanted domain. New reward function has been designed incorporating x and y. Generated moelcules have been screened using QSPR methods. Lastly, reactivity properties of top cnadidate molecules have been validated using DFT.

# Methods

## Data Preparation

Coconut is a large and freely available dataset of organic bio-based molecules. We selected 3500 molecules by restricting it to only C8-C16 saturated carbohydrates with up to one O atom. In addition, the dataset was supplemented with aviation fuel molecules from x. Molecules were represented as SMILES.

## TenGAN

Talk about Variant SMILES, Transformer Encoder as the discriminator and generator, RL, mini-batch discriminator.

### Reinforcement learning

Include full new reward function equation!

The objective function for the generator is given as:

Where denotes the action-value function, ….

Here, and are the property scores and the ratio of the number of unique SMILES strings to the product of the number of strings and the number of repeated strings, respectively. is a hyperparameter that is used to balance the strengths of GAN and RL.

### Hyperparameter Optimisation

Talk about initial HPE, list hyperparameter and ranges tested, test method. Talk about the effect of WGAN and Mini-batch. Talk about effect of pre-training.

### Comparison With Other Models

Big nice table with compatisons to other models, e.g. MolGAN, MolGPT, some VAE

## Molecule Screening

Talk about how molecules were screened and filtered.

## Validation

Talk about validation done using Gaussian 16, specifying the proccess and referencing the methodology from another paper.

# Results

Strctures, energies, orbitals, etc from Gaussian, Everything from tengan  
Heatmaps, loss curves, SHAP, feature importance, domain of applicability, bear plot, etc

![](data:application/pdf;base64,)

Example of evolution of TenGAN metrics over epochs.

![](data:application/pdf;base64,)

Heat map of performance metrics from hyperparameter optimisation.

![](data:application/pdf;base64,)

Effect of Wasserstein Distance and Mini-Batch discrimination on the output evaluation metrics.

# Discussion

internal and external. internal - compare model to itself; external - compare my results to what is SOTA.

# Conclusion

answer to the problem that we were trying to solve. open for perspectives.

# Outline

## Floats

New float types are automatically set up by the class file. The means graphics are included as follows (Scheme [[sch:example]](#sch:example)). As illustrated, the float is “here” if possible.

Your scheme graphic would go here: .eps format  
for LaTeX  or .pdf (or .png) for pdfLaTeX  
ChemDraw files are best saved as .eps files:  
these can be scaled without loss of quality, and can be  
converted to .pdf files easily using eps2pdf.

|  |
| --- |
| As well as the standard float types table and figure, the class also recognises scheme, chart and graph. |

An example figure

## Maths

Some inline material or followed by some display.

It is possible to label equations in the usual way (Eq. [[eqn:example]](#eqn:example)).

## Tables

An example table

| Header one | Header two |
| --- | --- |
| Entry one | Entry two |
| Entry three | Entry four |
| Entry five | Entry five |
| Entry seven | Entry eight |

The author thanks Dr. Zied Hosni (Institute of Materials Discovery, University College London, UK) for supervision and guidance. The author acknowledges the use of the UCL Myriad High Performance Computing Facility (Myriad@UCL), and associated support services, in the completion of this work.

Model files, GAN and Gaussian 16 log files, as well as scripts for figures available on:  
<https://github.com/ZiedHosni1/ZH5_Karl>  
Should include all log files from gaussian, orbitals, python files, etcetc. Some will selectively go into the main text.