香港中文大學
The Chinese University of Hong Kong

Leveraging large language models for predictive chemistry

Ansh Arora

30th May 2025

Nature Mach. Intell., 2024, 6, 161-169

Outline

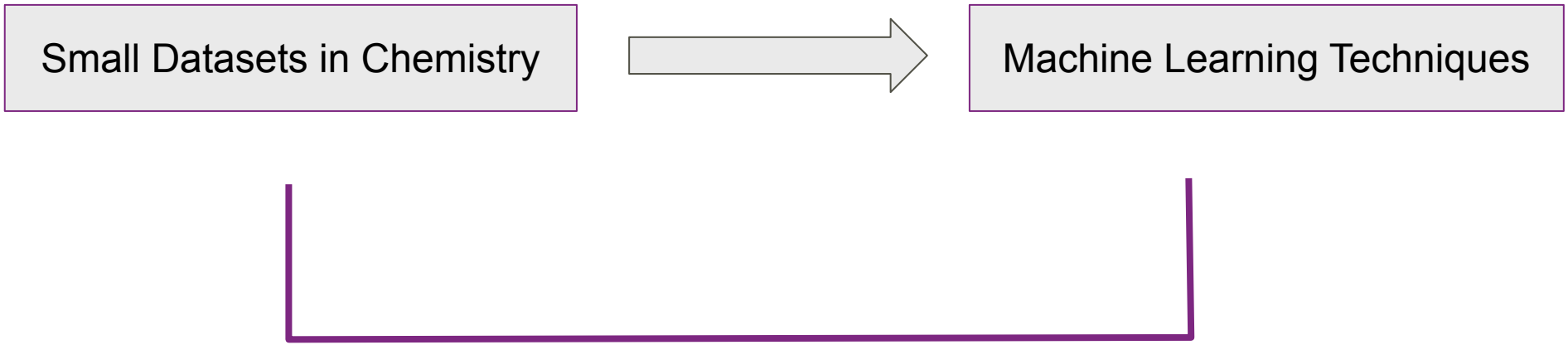
1. Background
2. Methodology
3. Applications
4. Conclusion
5. Future Outlook

Outline

Research Background

Research Background

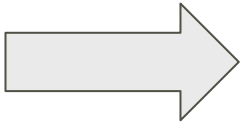
- Machine learning has transformed many fields and has recently found applications in chemistry and materials science.



- Perform these tasks with GPT-3 (a LLM)

Research Background

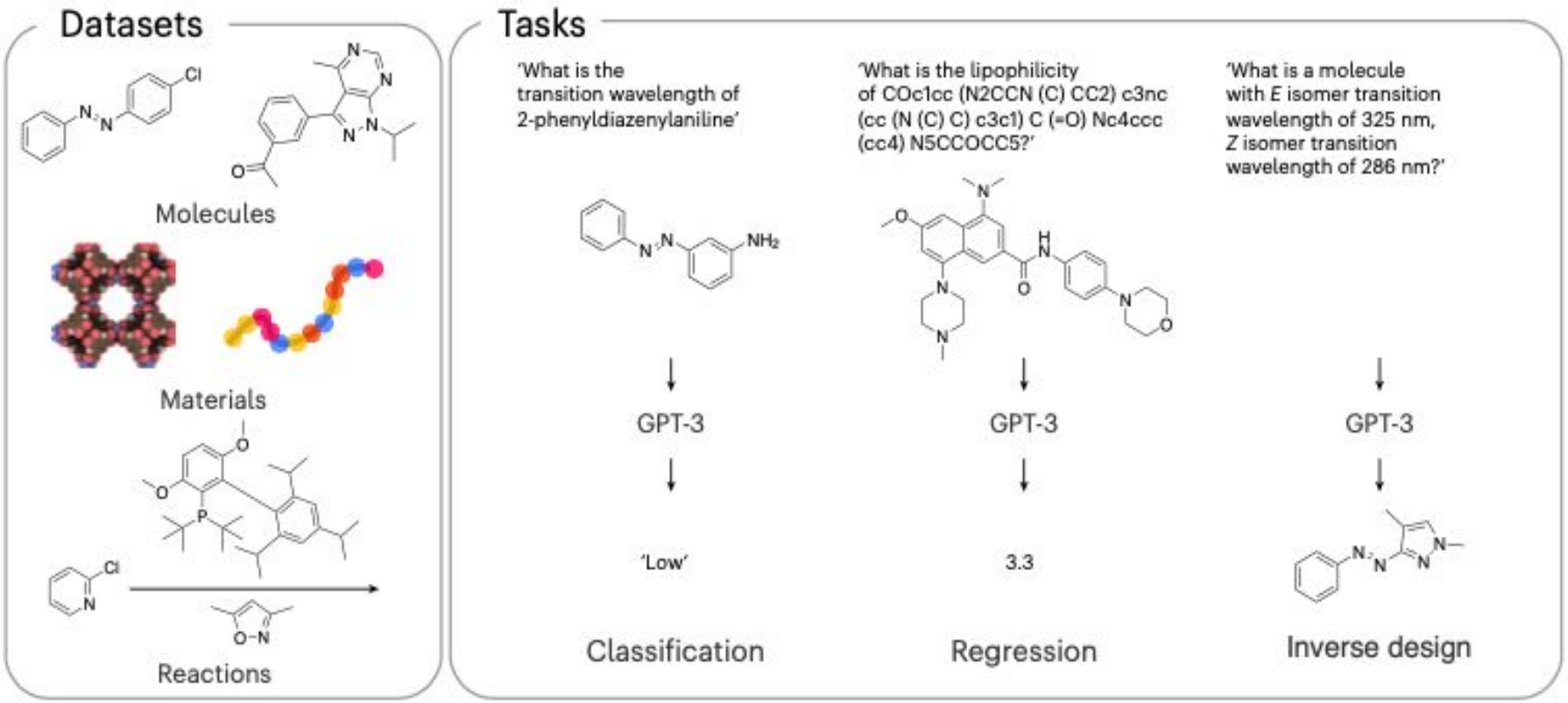
- But How?



We fine-tune it to answer chemical questions in natural language with the correct answer.

- Surprisingly, our fine-tuned version of GPT-3 performs comparably to or even outperforms conventional machine learning techniques, in particular low-data limit.

Research Background



❖ Fig. | Overview illustration of the datasets and tasks addressed in this work.

Outline

Methodology

Methodology

- Let us first discuss how we fine-tune the GPT-3 model for example, in the case of high-entropy alloys.
- The question we would like to ask is: 'What is the phase of <composition of the high-entropy alloy>?' and our model should give a text completion from the set of possible answers {single phase, multi-phase}.

In the Data Table, we can see the set of questions and answers we used to fine-tune the GPT-3 model.

prompt	completion	experimental
What is the phase of ColCulFe1Ni1V1?	0	multi-phase
What is the phase of Pu0.75Zr0.25?	1	single-phase
What is the phase of BeFe?	0	multi-phase
What is the phase of LiTa?	0	multi-phase
What is the phase of NbO.5Ta0.5?	1	single-phase
What is the phase of Al0.1W0.9?	1	single-phase
What is the phase of CrO.5Fe0.5?	1	single-phase
What is the phase of Al1Co1Cr1Cu1Fe1Ni1Ti1?	0	multi-phase
What is the phase of Cu0.5Mn0.5?	1	single-phase
What is the phase of OsU?	0	multi-phase

Methodology

- We selected this example to directly compare its performance with the current state-the-art machine learning models.

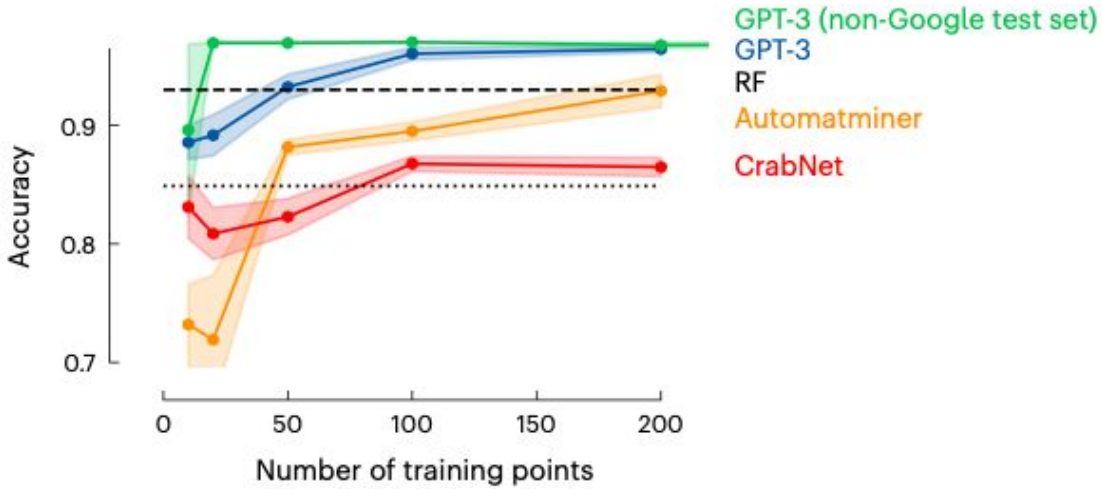


Fig. Accuracy of our GPT-3 model

The dashed horizontal line indicates the performance of the commonly used machine learning model using random forest (RF) with a dataset of 1,252 points and 10-fold cross-validation, that is, corresponding to a training set size of around 1,126 points.

We show that with only around 50 data points, we get a similar performance to the commonly used machine learning models, which were trained on more than 1,000 data points.

Outline

Applications

Applications

- We focused on those applications for which conventional machine learning methods have been developed and generally accepted as benchmarks in their field.

➔ Classification :

The following Data Table compares the performance of a fine-tuned GPT-3 model with baselines :

group	benchmark	publication year	best non-DL	best DL baseline
molecules	photoswitch transition wavelength	2022	1.1 (n)	1.2 (t)
	free energy of solvation	2014	3.1 (g)	1.3 (t)
	solubility	2004	1.0 (x)	0.002 (m)
	lipophilicity	2012	3.43 (g)	0.97 (t)
	HOMO-LUMO gap	2022	4.3 (x)	0.62 (t)
	OPV PCE	2018	0.95 (n)	0.76 (t)
materials	surfactant free energy of adsorption	2021	1.4 (xj)	0.37 (t)
	CO_2 Henry coefficients	2020	0.40 (x)	12 (t)
	CH_4 Henry coefficients	2020	0.52 (xmo)	0.60 (t)
	heat capacity	2022	0.24 (mo)	0.76 (c)
	HEA phase	2020	24 (prf)	9.0 (c)
	bulk metallic glass formation ability	2006	0.98 (a)	0.62 (mod)
	metallic behavior	2018	0.52 (a)	0.46 (mod)
reactions	C-N cross-coupling	2018	2.9 (drfp)	
	C-C cross-coupling	2022	0.98 (n)	

For the analysis in this table, we fit the learning curves for the GPT-3 models and for the baselines and measure where the learning curves intersect.

We determine the factor of how much more (or less) data we would need to make the best baseline perform equal to the GPT-3 models in the low-data regime of the learning curves.

Applications

- We studied molecular properties like HOMO–LUMO gaps, water solubility, and photovoltaic performance; material properties of alloys, MOFs, and polymers; and two key cross-coupling reactions in organic chemistry.

A learning curve, for example in the case of free energy of solvation is shown.

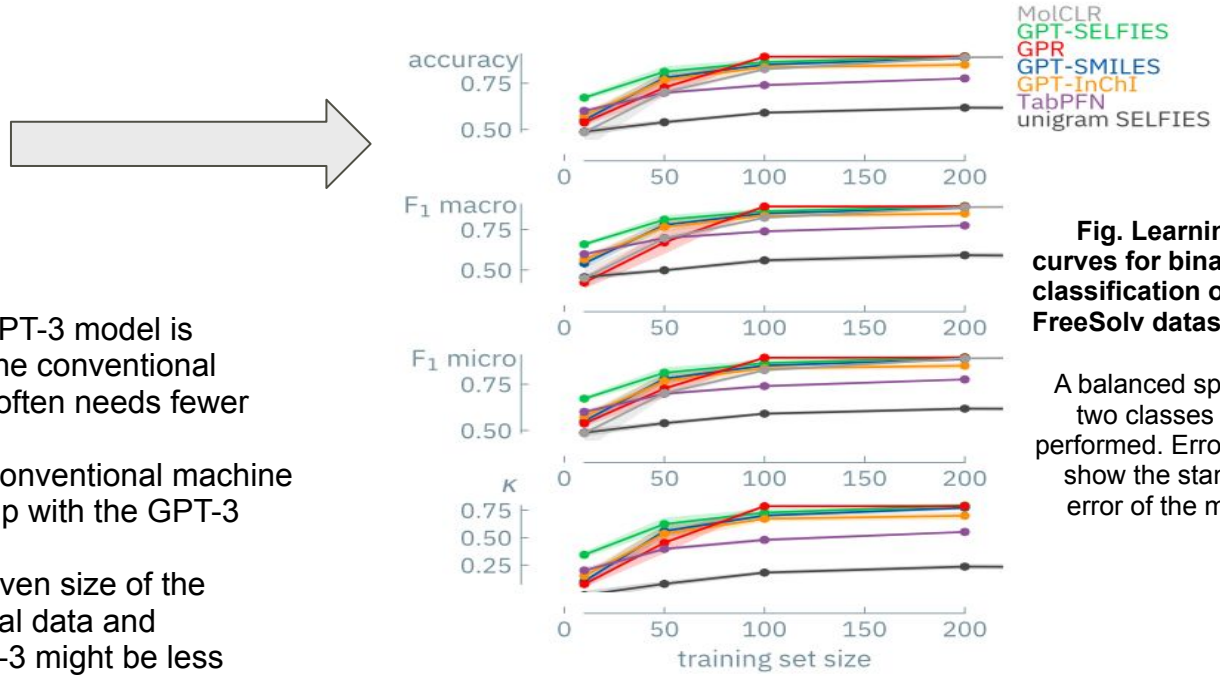


Fig. Learning curves for binary classification on the FreeSolv dataset

A balanced split into two classes was performed. Error bands show the standard error of the mean.

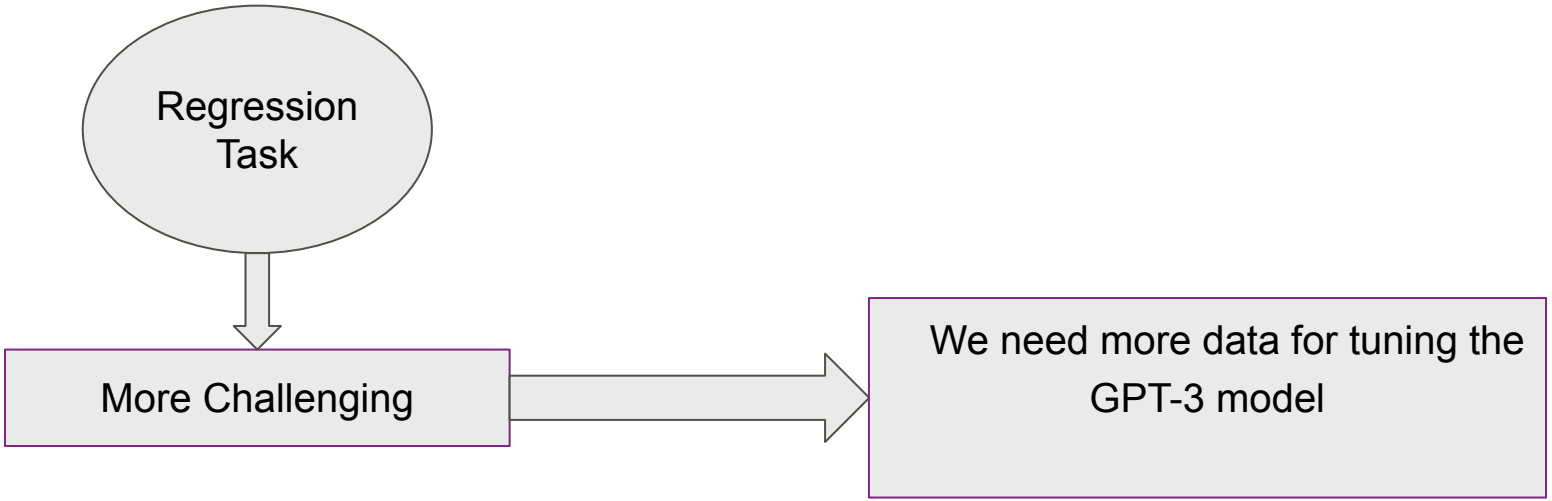
- In the low-data regime, our GPT-3 model is typically at least as good as the conventional machine learning model and often needs fewer data.
- In the high-data regime, the conventional machine learning models often catch up with the GPT-3 model.
- This makes sense, as for a given size of the dataset, the need for additional data and correlations captured by GPT-3 might be less needed.

Applications

➔ **Regression :**

- Would allow us to predict the value of a continuous property such as the Henry coefficient for the adsorption of a gas in a porous material.

- As we are using a pre-trained language model, performing actual regression that predicts real numbers ($\in$ R) is impossible (without changes to the model architecture and training procedure).



- We still get a performance that can approach the state of the art, but as this approach requires much more data, the advantage, except for the ease of training, is less.

Applications

- One Way to Approximate Regression: directly predicting rounded floating point numbers.
- One would expect the performance to be **worse** than in the classification setting. GPT-3 performs worse than baselines in this setting.
- However, it sometimes approaches the performance of the baselines.

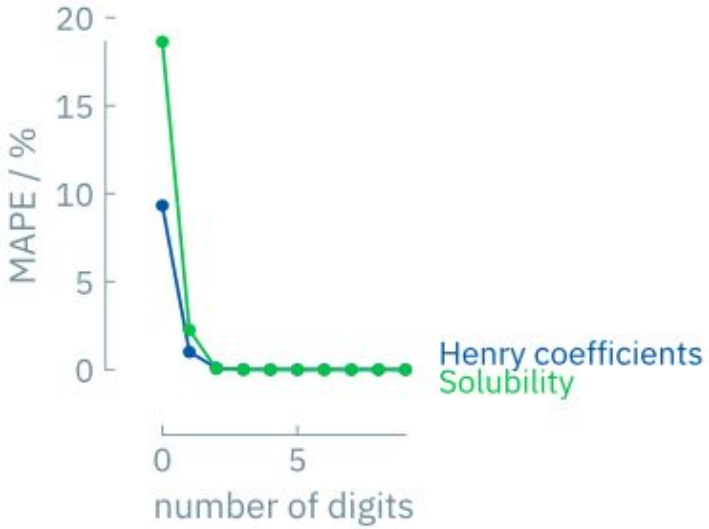


Figure: Mean absolute percentage error due to rounding to different numbers of digits on two datasets.

→ **Errors due to rounding**

Applications

➔ Inverse Design :

- Generating molecules with the desired set of properties.

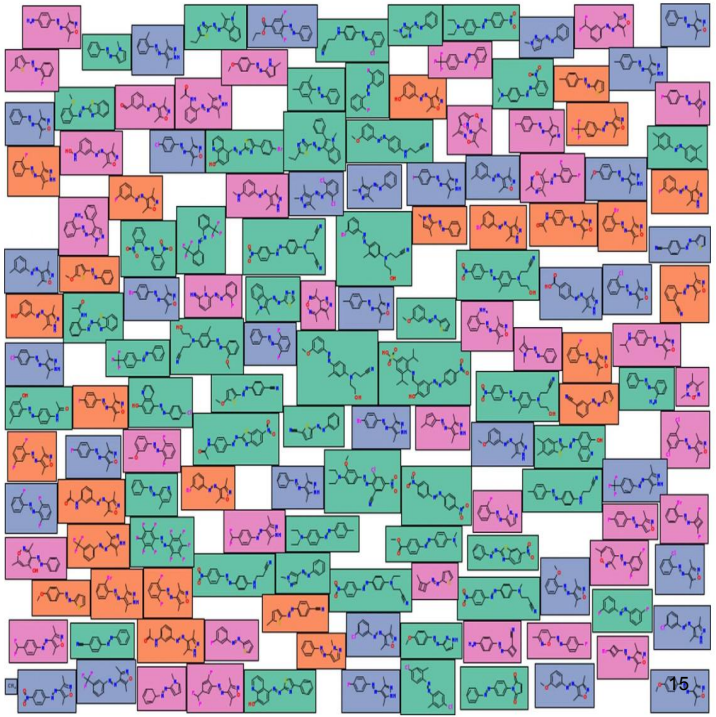
- With limited dataset, GPT-3 can still predict properties effectively, making it ideal for early-stage research. We could leverage a fine-tuned GPT-3 model to generate suggestions for novel materials with similar or potentially improved performance through inverse design.

- For GPT-3, inverse design is as simple as training the model with question and completion reversed. That is, answer the question ‘What is a photoswitch with transition wavelengths of 324 nm and 442 nm, respectively’ with a text completion that should be a string of a meaningful molecule.

- Test to see whether our model can generate new structures upon asking it to generate molecules with transition wavelengths similar to those from the provided dataset :

Extended Data Fig. 1: Molecule Cloud for randomly generated photoswitch molecules.

From: Leveraging large language models for predictive chemistry



Applications

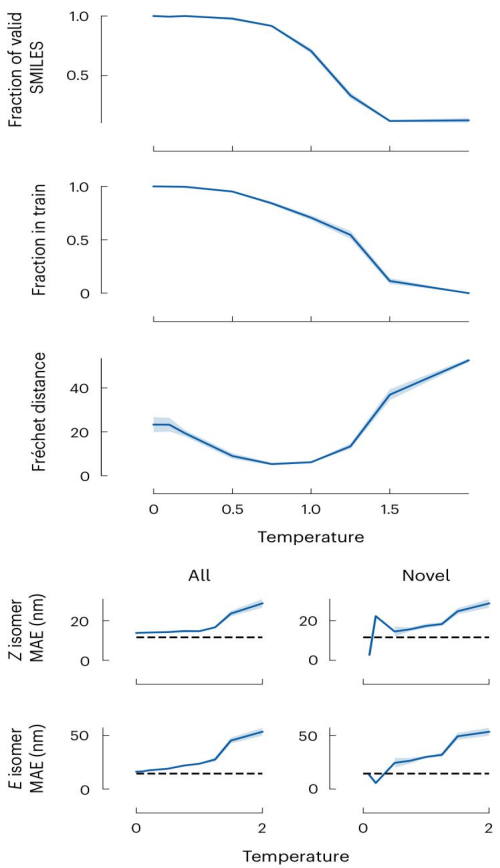


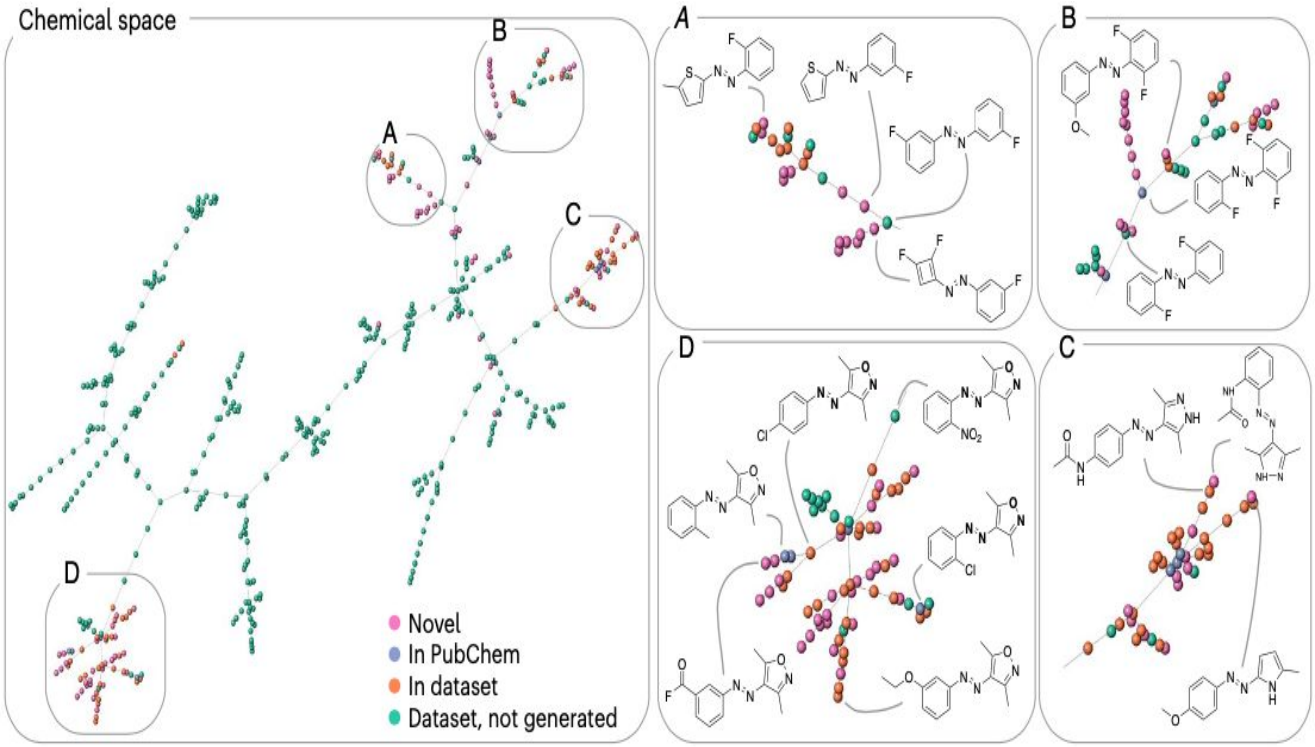
Fig. 3 | Photoswitch inverse design metrics as a function of temperature.

- The fraction of valid SMILES indicates the fraction of generated SMILES that can successfully be parsed using RDKit.
- We then determine the fraction of those that are already part of the training set and find that at low temperature GPT-3 tends to restate molecules from the training set.
- To quantitatively capture the similarity of the distribution of the generated molecules to the ones from the training set, we compute the Fréchet ChemNet distance, which quantifies both diversity and distribution match and goes through a minimum at intermediate temperatures.
- For quantifying how well the generated molecules match the desired transition wavelengths, we use the GPR models reported by ref. 43 to predict the transition wavelengths. The dashed horizontal lines indicate those models' mean absolute error (MAE). Across all temperatures, we found high average synthesizability.

Applications

- To assess novelty, we compared our generated molecules to those in ref. 43 using molecular fingerprint distances. Figure 4 visualizes this by laying out the resulting approximate nearest-neighbour graph in two dimensions.

Fig. 4 | TMAP visualization of the generated photoswitches and the training set.



From this figure, we see that the generated molecules sometimes act as substitutions for existing ones in the dataset, while in other cases, they introduce entirely new scaffolds.

Outline

Conclusion

Conclusion

Our results raise a very important question: how can a natural language model with no prior training in chemistry outperform dedicated machine learning models, as we were able to show in the case of high-entropy alloys and for various molecule, material and chemical reaction properties?

- To our knowledge, this fundamental question has no rigorous answer. The fact that we get good results independent of the chemical representation illustrates that these language models are very apt at extracting correlations from any text.
- A machine learning system built using GPT-3 works impressively well for a wide range of questions in chemistry—even for those for which we cannot use conventional line representations such as SMILES. Compared with conventional machine learning, it has many advantages.
- GPT-3 supports diverse applications using a unified natural language Q&A approach, setting a strong baseline that future ML models must surpass.
- Using GPT-3 in research is like performing a literature search—it uncovers meaningful chemical correlations in text, offering new opportunities for chemists and materials scientists.

Outline

Future Outlook

Future Outlook

- If we say that the GPT-3 model is successful, it implies only that we have established that the GPT-3 model has identified correlations in the current training data that can be successfully exploited to make predictions. However, this does not imply that the correlations are always meaningful or related to cause–effect relationships.

- Hence, our research does not stop here. The next step will be to use GPT-3 to identify these correlations and ultimately get a deeper understanding.

- We argue that GPT-3 is only a tool to make more effective use of the knowledge scientists have collected over the years.

- It is also important to mention that while the training corpus contains chemistry information, many, if not most, scientific articles and results have not been seen by GPT-3. Hence, one can expect an even more impressive performance if these data are added to the training data.

Questions?

Thank You