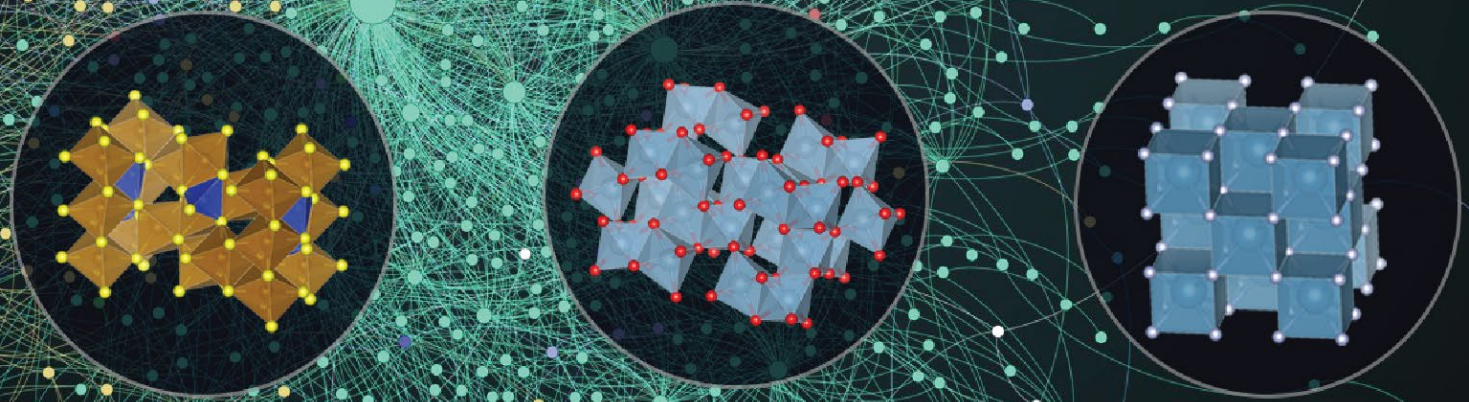
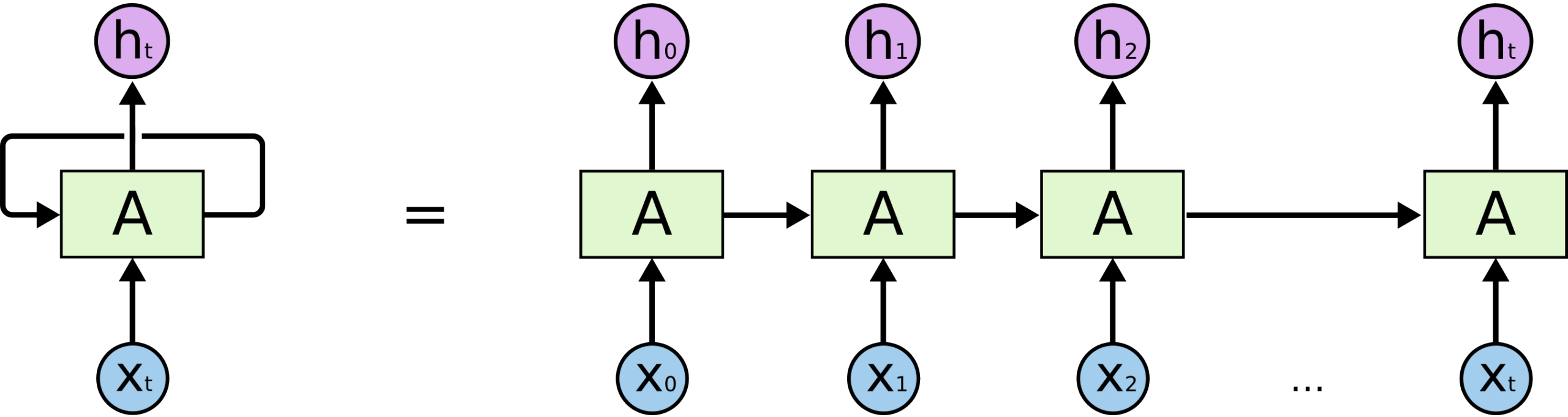
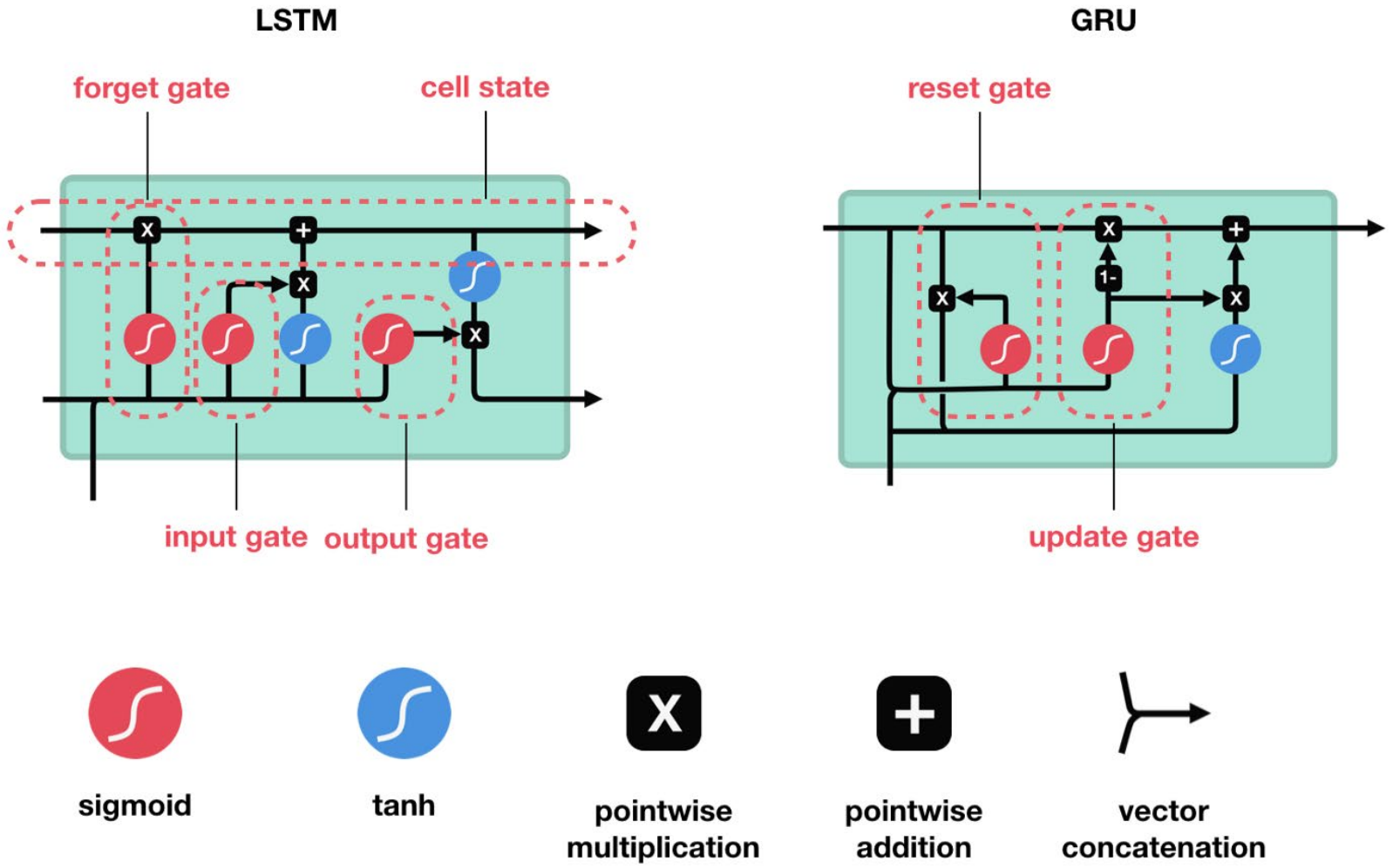
Transformers



Transformers were created to fix some of the problems of other recurrent neural networks



LSTMs were an improvement on RNNs, but they were too slow and still too limited



Michael Phi, Illustrated Guide to LSTM's and GRU's: A step by step explanation, YouTube

Transformers allow us to parallelize sequential data!

Attention Is All You Need

Ashish Vaswani* Google Brain avaswani@google.com
Noam Shazeer* Google Brain noam@google.com
Niki Parmar* Google Research nikip@google.com
Jakob Uszkoreit* Google Research usz@google.com
Llion Jones* Google Research llion@google.com
Aidan N. Gomez* † University of Toronto aidan@cs.toronto.edu
Łukasz Kaiser* Google Brain lukaszkaiser@google.com
Illia Polosukhin* ‡ illia.polosukhin@gmail.com

Abstract

The dominant sequence transduction models are based on complex recurrent or convolutional neural networks that include an encoder and a decoder. The best performing models also connect the encoder and decoder through an attention mechanism. We propose a new simple network architecture, the Transformer, based solely on attention mechanisms, dispensing with recurrence and convolutions entirely. Experiments on two machine translation tasks show these models to be superior in quality while being more parallelizable and requiring significantly less time to train. Our model achieves 28.4 BLEU on the WMT 2014 English-to-German translation task, improving over the existing best results, including ensembles, by over 2 BLEU. On the WMT 2014 English-to-French translation task, our model establishes a new single-model state-of-the-art BLEU score of 41.8 after training for 3.5 days on eight GPUs, a small fraction of the training costs of the best models from the literature. We show that the Transformer generalizes well to other tasks by applying it successfully to English constituency parsing both with large and limited training data.

1 Introduction

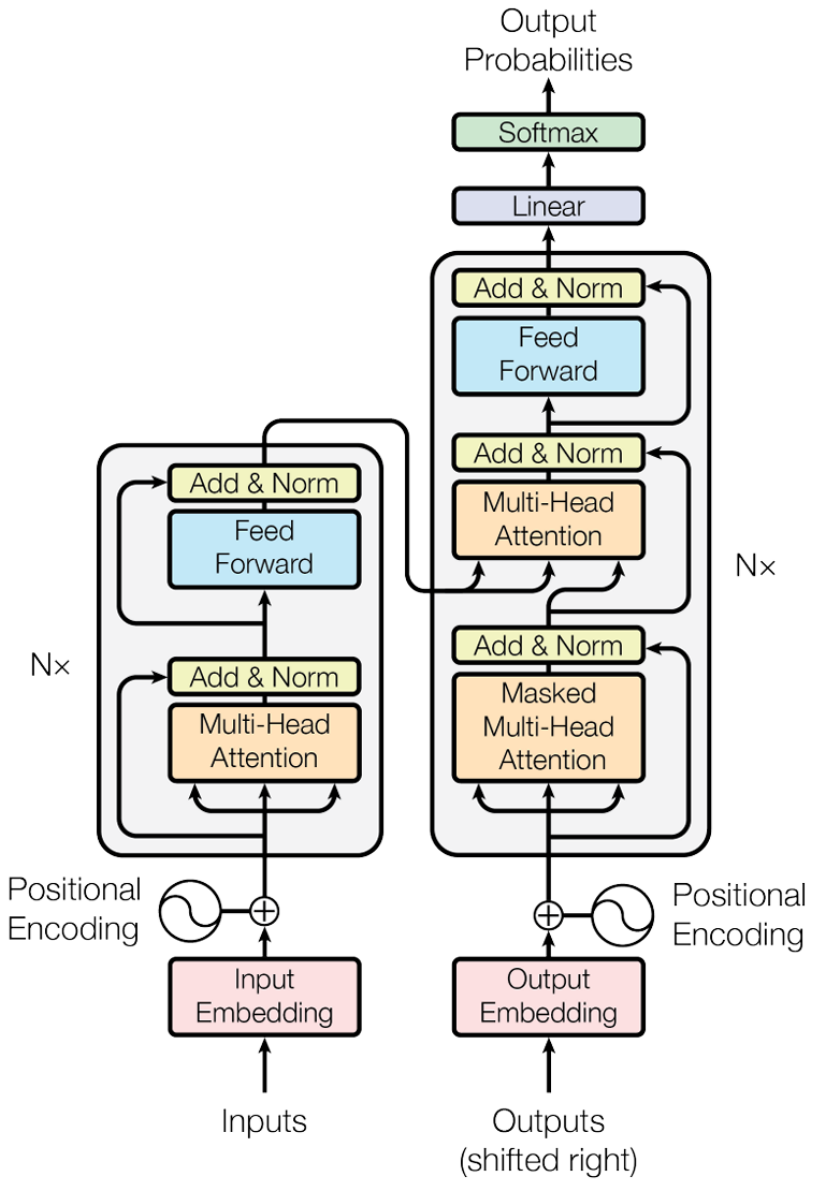
Recurrent neural networks, long short-term memory [13] and gated recurrent [7] neural networks in particular, have been firmly established as state of the art approaches in sequence modeling and

*Equal contribution. Listing order is random. Jakob proposed replacing RNNs with self-attention and started the effort to evaluate this idea. Ashish, with Illia, designed and implemented the first Transformer models and has been crucially involved in every aspect of this work. Noam proposed scaled dot-product attention, multi-head attention and the parameter-free position representation and became the other person involved in nearly every detail. Niki designed, implemented, tuned and evaluated countless model variants in our original codebase and tensor2tensor. Llion also experimented with novel model variants, was responsible for our initial codebase, and efficient inference and visualizations. Lukasz and Aidan spent countless long days designing various parts of and implementing tensor2tensor, replacing our earlier codebase, greatly improving results and massively accelerating our research.

†Work performed while at Google Brain.

‡Work performed while at Google Research.

31st Conference on Neural Information Processing Systems



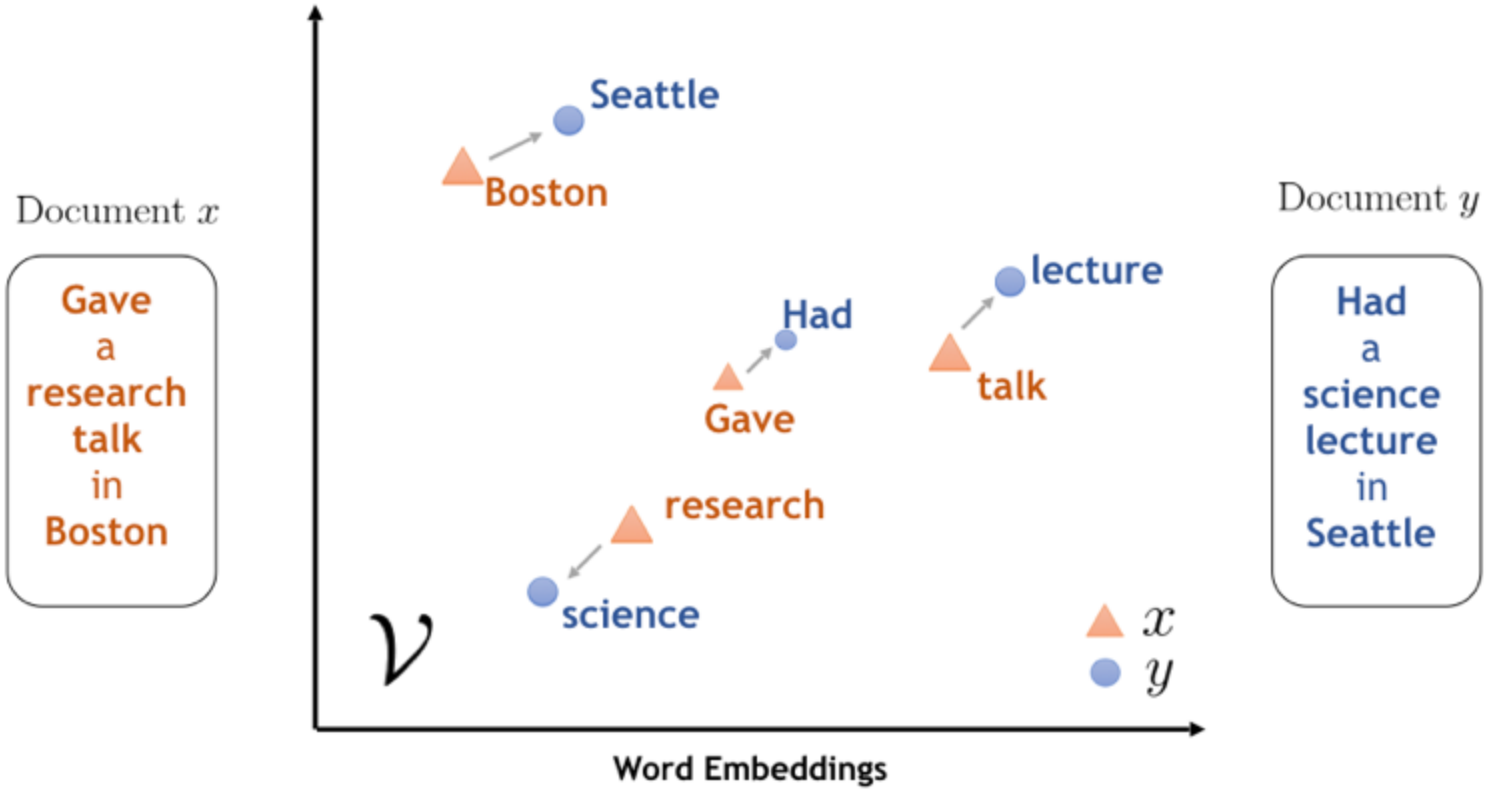
[CITATION] **Attention** is **all you need**

A Vaswani - Advances in Neural Information Processing Systems, 2017

☆ Save 99 Cite Cited by 153082 Related articles

Vaswani et al., Attention is all you need, NIPS (2017)

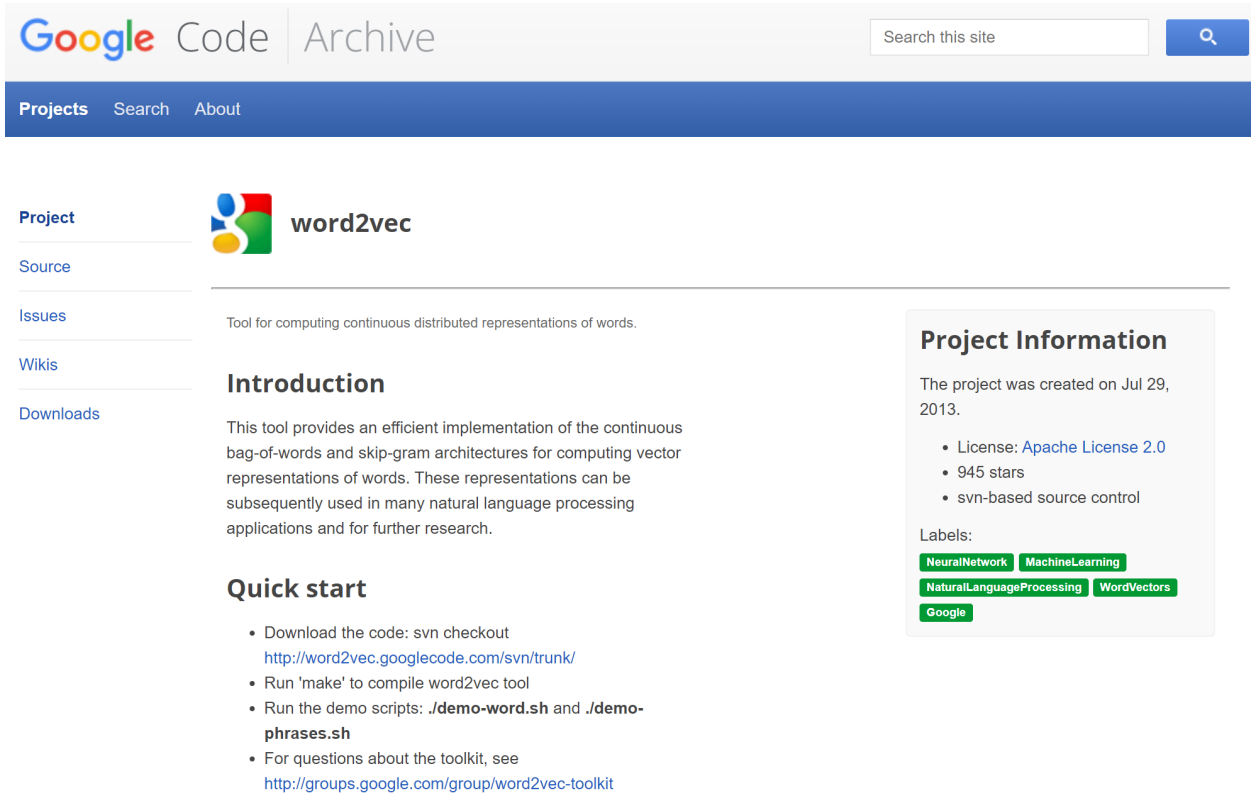
Transformers first use embeddings to featurize the input*

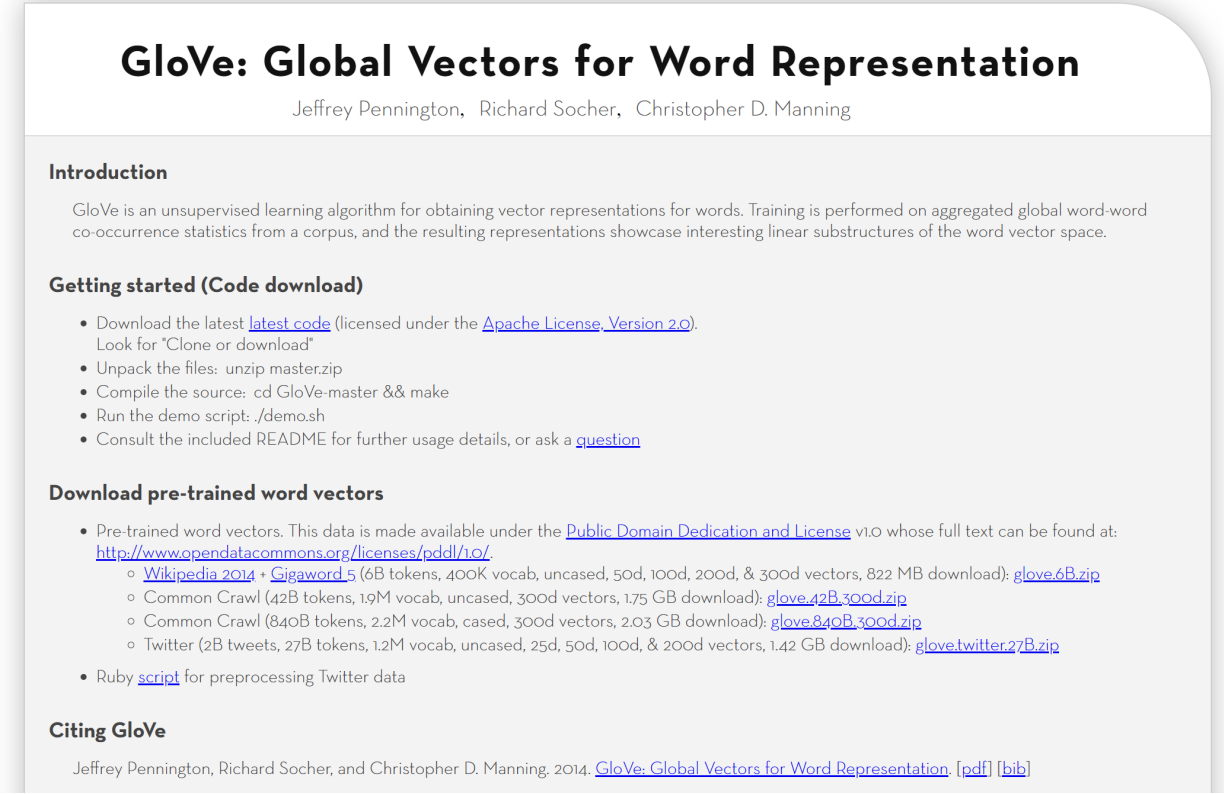


*needed for words, but could be skipped for some inputs

https://www.ibm.com/blogs/research/2018/11/word-movers-embedding/

Embeddings can be generated from training data, or rely on pretrained embeddings





https://code.google.com/archive/p/word2vec/ & https://nlp.stanford.edu/projects/glove/

Inputs can have different meanings depending on sequence and context

The deadline creeped up on me.
The tensile specimen creeped until failure

She conducts the music beautifully.
The ceramic only conducts ions.

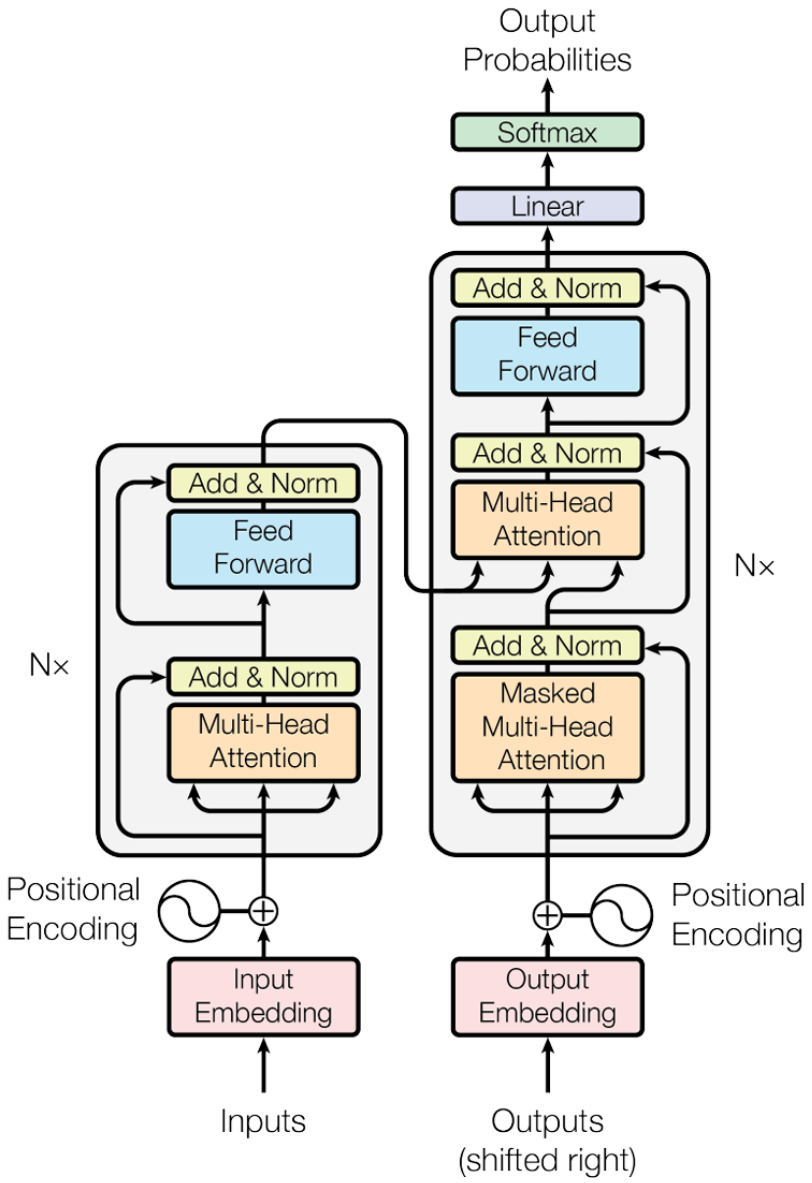
I observed necking in my sample.
I observed necking in the parked cars at lover's lookout.

To get the right meaning, we need to use a positional encoding

$$\begin{bmatrix} 0.56 \\ 0.77 \\ 0.1 \end{bmatrix} + positional\ encoding = embedding\ in\ context$$

$$PE = \cos(position/1000^{2i/d\_model})$$

Inputs are passed to "encoder block" with multi-head attention and feed forward layers



Vaswani et al., Attention is all you need, NIPS (2017)

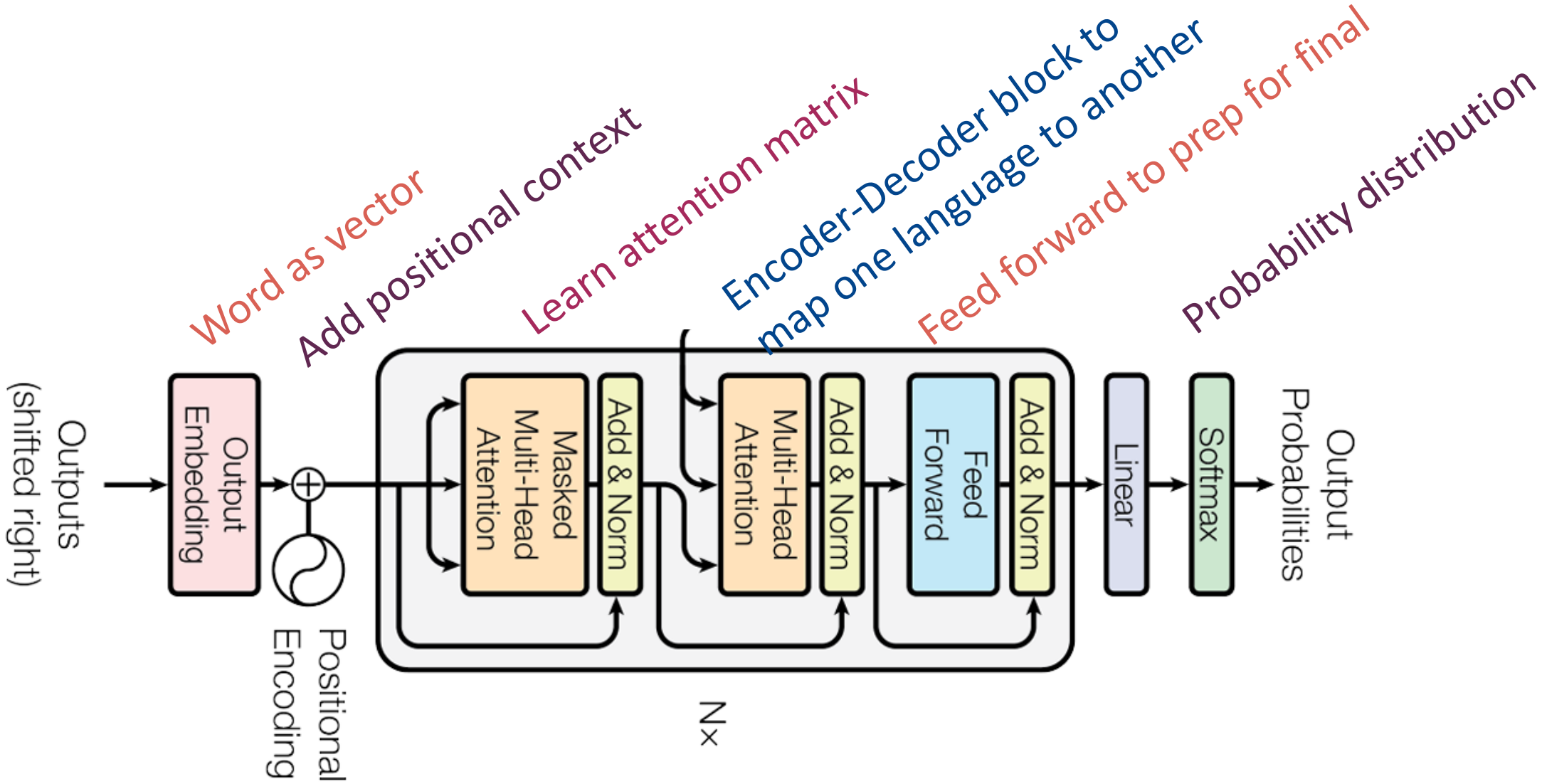
Attention vectors tell us what each input pays attention to in the sequence of data

The ceramic only conducts ions.

$$\begin{bmatrix} The \\ ceramic \\ only \\ conducts \\ ions \end{bmatrix} \xrightarrow{focus} \begin{bmatrix} The & ceramic & only & conducts & ions \\ The & ceramic & only & conducts & ions \\ The & ceramic & only & conducts & ions \\ The & ceramic & only & conducts & ions \\ The & ceramic & only & conducts & ions \end{bmatrix}$$

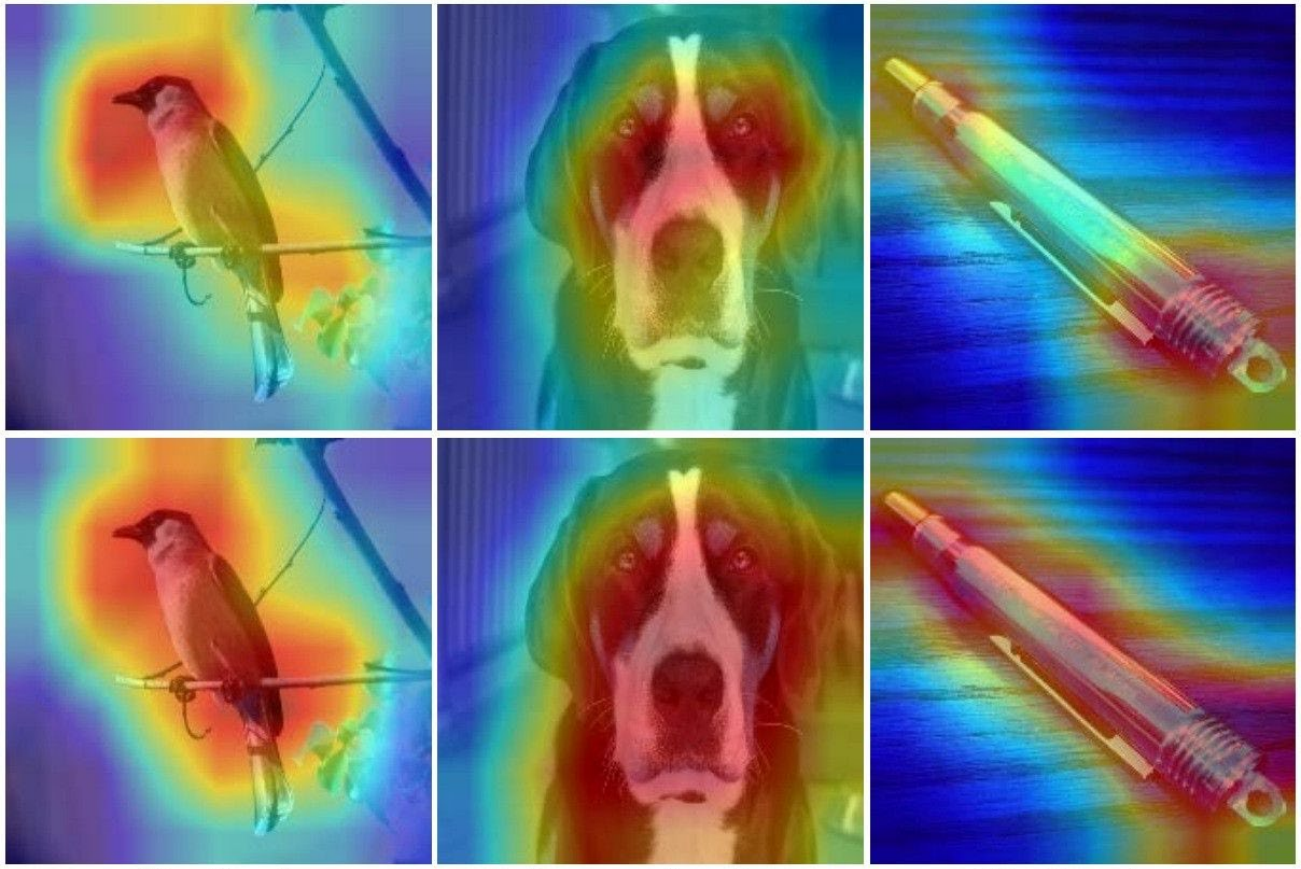
These attention vectors for each input are then sent through feed forward layer to transform these vectors into a layer that can be used for next step: “the decoder block”

The decoder block generates output sequence one element at a time based on encoder input



CodeEmporium, Transformer Neural Networks - EXPLAINED! (Attention is all you need), YouTube

Multi-head attention is used to get multiple perspectives on where to focus

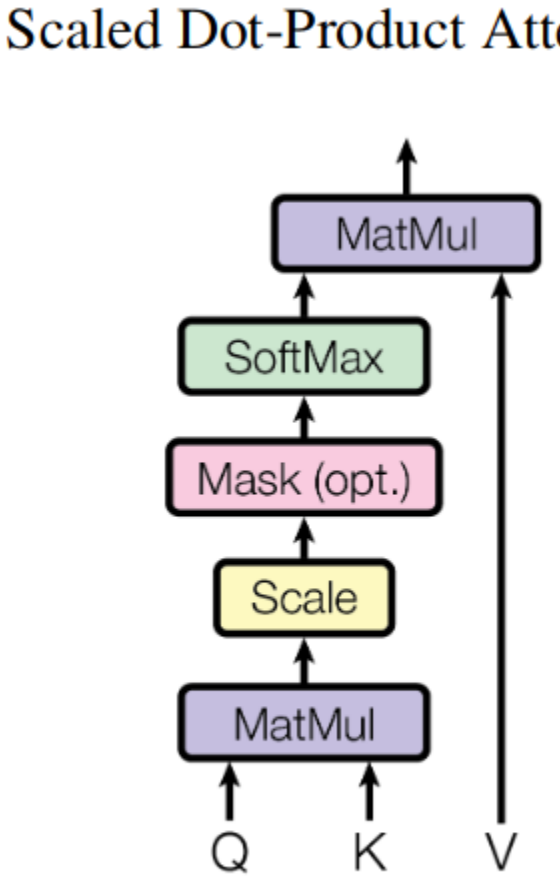


BONUS: each attention map is done independently from others, so we can parallelize this part of the process!

Encode entire sequence in parallel

https://blog.paperspace.com/attention-mechanisms-in-computer-vision-cbam/

Each attention layer contains query, key, and value to learn relationships

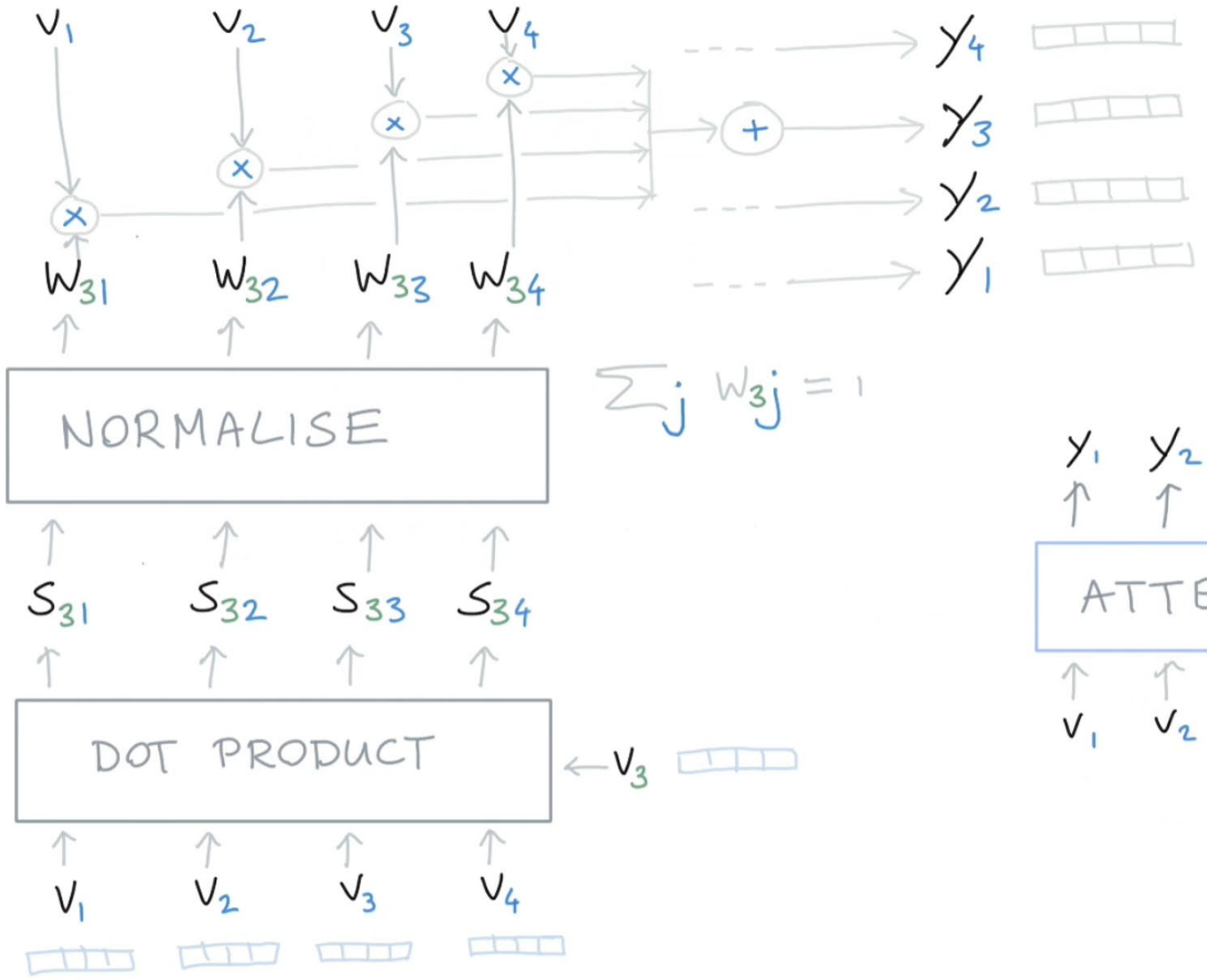


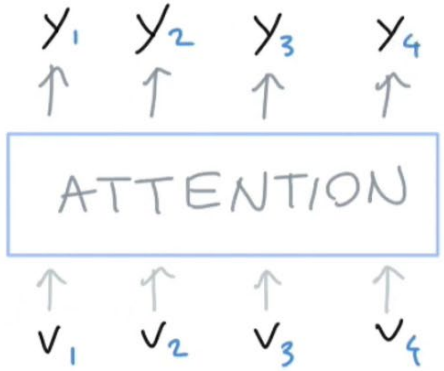
We use Q, K, V (abstract vectors) to compute an attention vector for each word

$$Z = softmax\left(\frac{Q.K^T}{\sqrt{Dimension\ of\ vector\ Q, K, or\ V}}\right).V$$

Vaswani et al., Attention is all you need, NIPS (2017)

The attention block adds context!

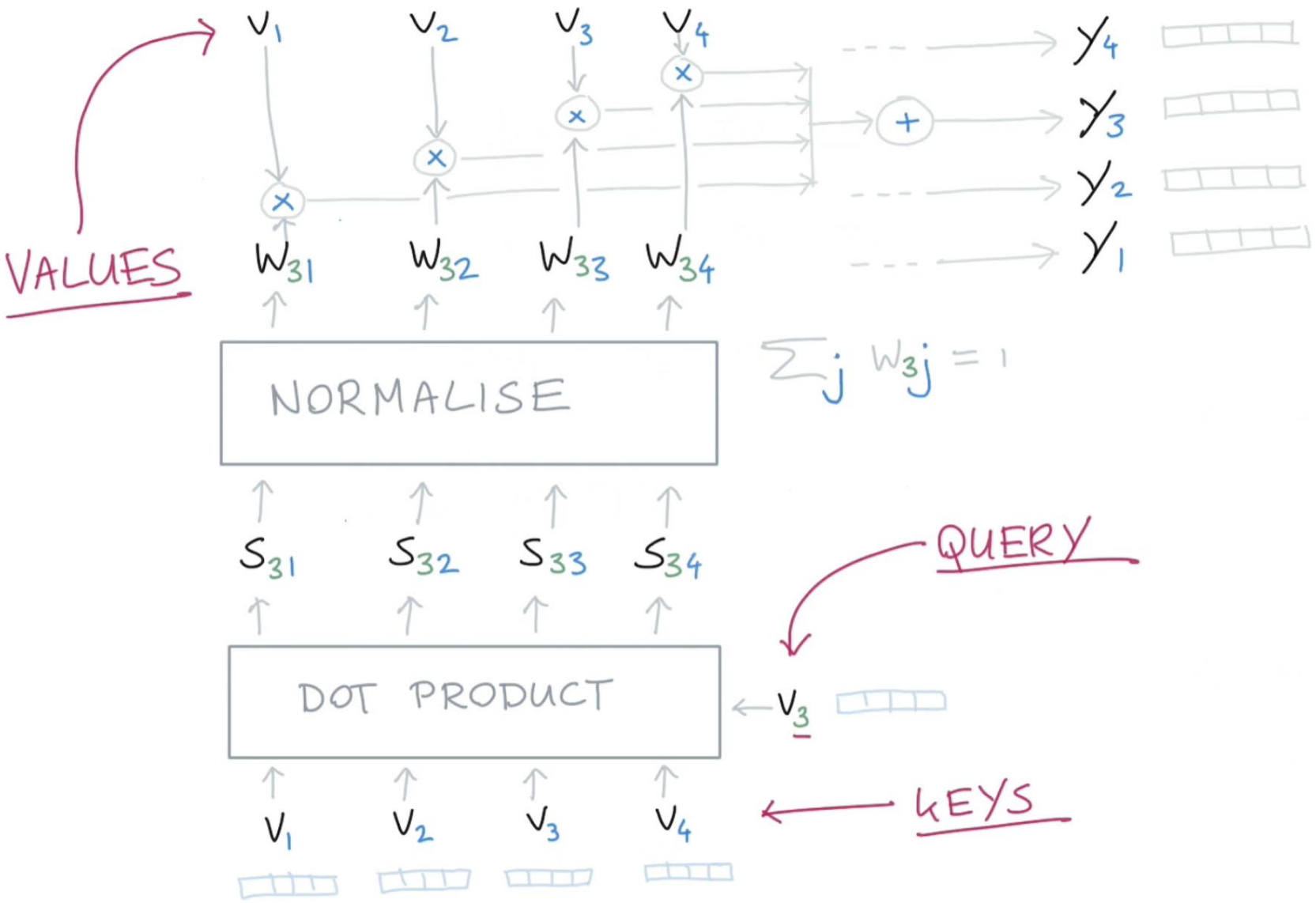




None of this has any tunable parameters to learn during training!

Rasa, Rasa Algorithm Whiteboard - Transformers & Attention 2: Keys, Values, Queries, YouTube

We can add tunable parameters to make context pattern finding more efficient



Keys: $v_i M_k$

Queries: $v_i M_Q$

Values: $v_i M_V$

Rasa, Rasa Algorithm Whiteboard - Transformers & Attention 2: Keys, Values, Queries, YouTube

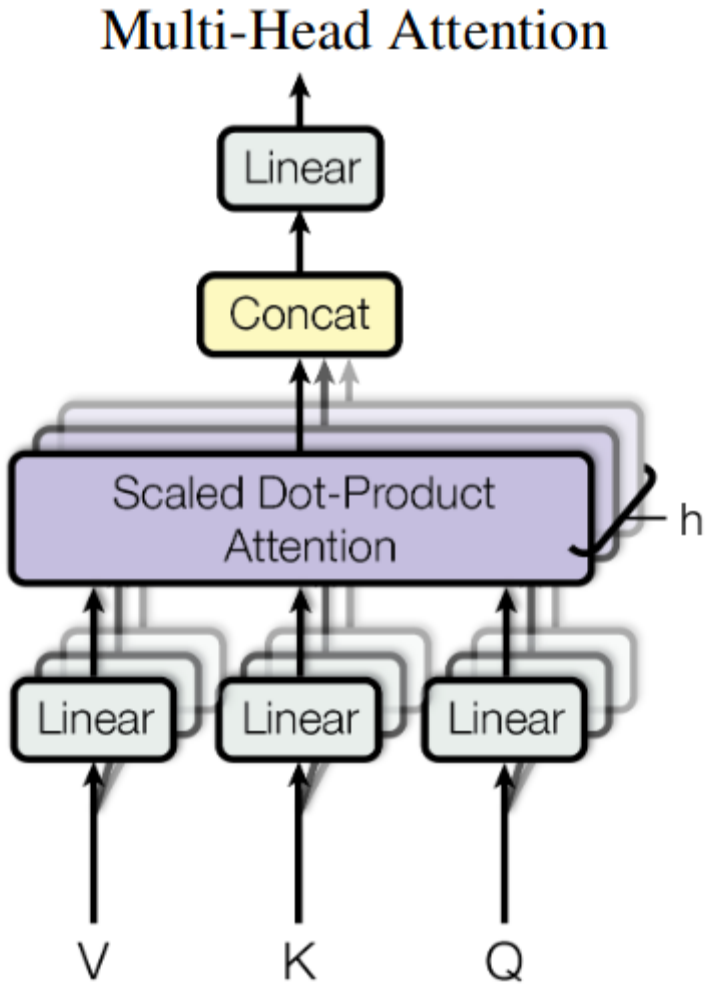
Multi-head attention requires multiple attention vectors for each input

We now have matrices of W^V, W^K, W^Q

$$Z = softmax\left(\frac{Q.K^T}{\sqrt{Dimension\ of\ vector\ Q, K, or\ V}}\right).V$$

Since neural net only expects one vector per word, we flatten to a single vector per word with a weighted matrix

$$\mathrm{Z} = [\mathrm{Z}_1, \mathrm{Z}_2, \mathrm{Z}_3 \ldots].W^Z$$



Vaswani et al., Attention is all you need, NIPS (2017)

Tensorflow has a nice code repository for attention networks

Neural machine translation with attention

Run in Google Colab | View source on GitHub | Download notebook

This notebook trains a sequence to sequence (seq2seq) model for Spanish to English translation based on Effective Approaches to Attention-based Neural Machine Translation. This is an advanced example that assumes some knowledge of:

- Sequence to sequence models
- TensorFlow fundamentals below the keras layer:
  - Working with tensors directly
  - Writing custom `keras.Model`s and `keras.layers`

While this architecture is somewhat outdated it is still a very useful project to work through to get a deeper understanding of attention mechanisms (before going on to Transformers).

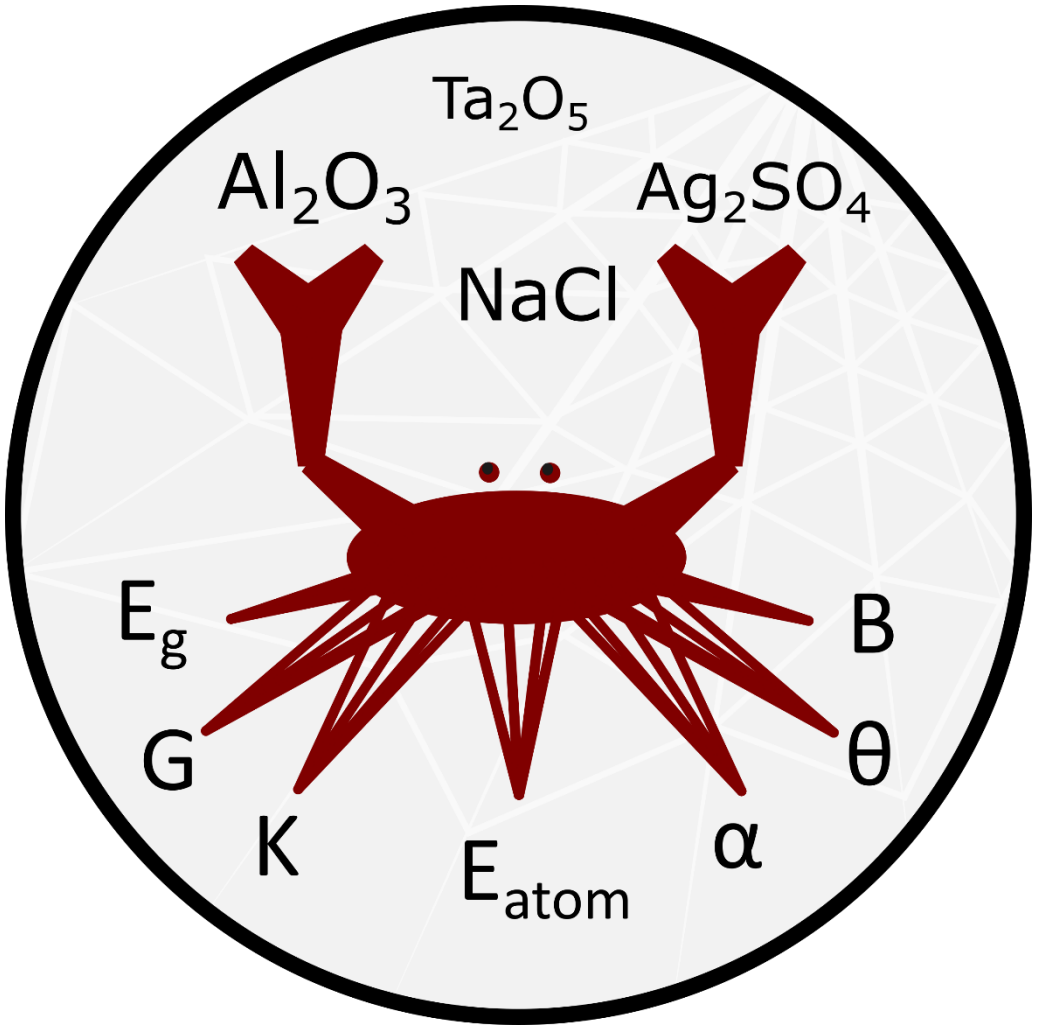
After training the model in this notebook, you will be able to input a Spanish sentence, such as "*¿todavia estan en casa?*", and return the English translation: "*are you still at home?*"

The resulting model is exportable as a `tf.saved_model`, so it can be used in other TensorFlow environments.

The translation quality is reasonable for a toy example, but the generated attention plot is perhaps more interesting. This shows which parts of the input sentence has the model's attention while translating:

https://www.tensorflow.org/text/tutorials/nmt_with_attention

Compositionally-restricted attention-based network (CrabNet)



npj Computational Materials

www.nature.com/npjcompumats

ARTICLE OPEN

Check for updates

Compositionally restricted attention-based network for materials property predictions

Anthony Yu-Tung Wang[1,3], Steven K. Kauwe[2,3], Ryan J. Murdock[2] and Taylor D. Sparks[2✉]

In this paper, we demonstrate an application of the Transformer self-attention mechanism in the context of materials science. Our network, the Compositionally Restricted Attention-Based network (CrabNet), explores the area of structure-agnostic materials property predictions when only a chemical formula is provided. Our results show that CrabNet's performance matches or exceeds current best-practice methods on nearly all of 28 total benchmark datasets. We also demonstrate how CrabNet's architecture lends itself towards model interpretability by showing different visualization approaches that are made possible by its design. We feel confident that CrabNet and its attention-based framework will be of keen interest to future materials informatics researchers.

npj Computational Materials (2021)7:77 ; https://doi.org/10.1038/s41524-021-00545-1

INTRODUCTION

Materials scientists constantly strive to achieve better understanding, and therefore better predictions, of materials properties. This began with the collection of empirical evidence through repeated experimentation, resulting in mathematical generalizations, theories, and laws. More recently, computational methods have arisen to solve a large variety of problems that were intractable to analytical approaches alone[1,2].

As experimental and computational methods have become more efficient, high-quality data has opened up a new avenue to materials understanding. Materials informatics (MI) is the resulting field of research that utilizes statistical and machine learning (ML) approaches in combination with high-throughput computation to analyze the wealth of existing materials information and gain unique insights[2–4]. As this wealth has increased, practitioners of MI have increasingly turned to deep learning techniques to model and represent inorganic chemistry, resulting in approaches such as ElemNet, IRNet, CGCNN, SchNet, and Roost[5–9]. In specific cases including CGCNN and SchNet, the compounds are represented using their chemical and structural information[7,8,10–15].

Modeling approaches based on crystal structure are an excellent tool for MI. Unfortunately, there are many material property datasets that lack suitable structural information. An example of this is the experimental band gap data gathered by Zhou et al.[16]. Conversely, many databases such as the Inorganic Crystal Structure Database (ICSD) and Pearson's Crystal Data (PCD) contain an abundance of structural information, but lack the associated material properties of the recorded structures. In both cases, the applicability of structure-based learning approaches are limited. This limitation is particularly evident in the discovery of novel materials, since it is not possible to know the structural information of (currently undiscovered) chemical compounds a priori. Therefore, the development of structure-agnostic techniques is well-suited to the discovery of novel materials.

A typical approach to structure-agnostic learning is done by representing chemistry as a composition-based feature vector (CBFV)[17]. This allows for data-driven learning in the absence of structural information. The CBFV is a common way to transform chemical compositions into usable features for ML and is generated from the descriptive statistics of a compound's constituent element properties. Researchers have effectively used CBFV-based ML techniques to generate material property predictions[17–25].

One potential issue with the CBFV approach lies in the way the element vectors are combined to form the vector describing the chemical compound. Typically, the individual element vectors of the compound are scaled by the element's prevalence (fractional abundance) in the composition, before being used to form the CBFV. This step assumes that the stoichiometric prevalence of constituent elements in a compound dictates their chemical signal, or contribution, to the material's property. However, this is not true in all cases; an extreme example of this is element doping. Dopants can be present in very small amounts in a compound, but can have a significant impact on its electrical[23,26,27], mechanical[20,28–30], and thermal properties[31–34]. In the case of a typical CBFV approach that uses the weighted average of element properties as a feature, the signal from dopant elements would not significantly change the vector representation of a compound. As a result, the trained ML model would fail to capture a portion of the relevant chemical information available in the full composition.

It is apparent that there is no generally accepted best way to model materials property behaviors. Different ML approaches lend themselves towards different modeling tasks. CGCNN requires access to structural information, ElemNet operates within the realm of large data, and classical models excel when domain knowledge can be exploited to overcome data scarcity[35]. To address the diversity of learning challenges, in Dunn et al., the Automatminer framework uses computationally expensive searches to optimize classical modeling techniques. They demonstrate effective learning on some data, but show shortcomings when deep learning is appropriate[36].

In a similar spirit, we seek to overcome general challenges in the area of structure-agnostic learning using an approach we refer to as the Compositionally Restricted Attention-Based network (CrabNet). CrabNet introduces the self-attention mechanism to the task of materials property predictions, and dynamically learns and updates individual element representations based on their chemical environment. To enable this, we introduce a

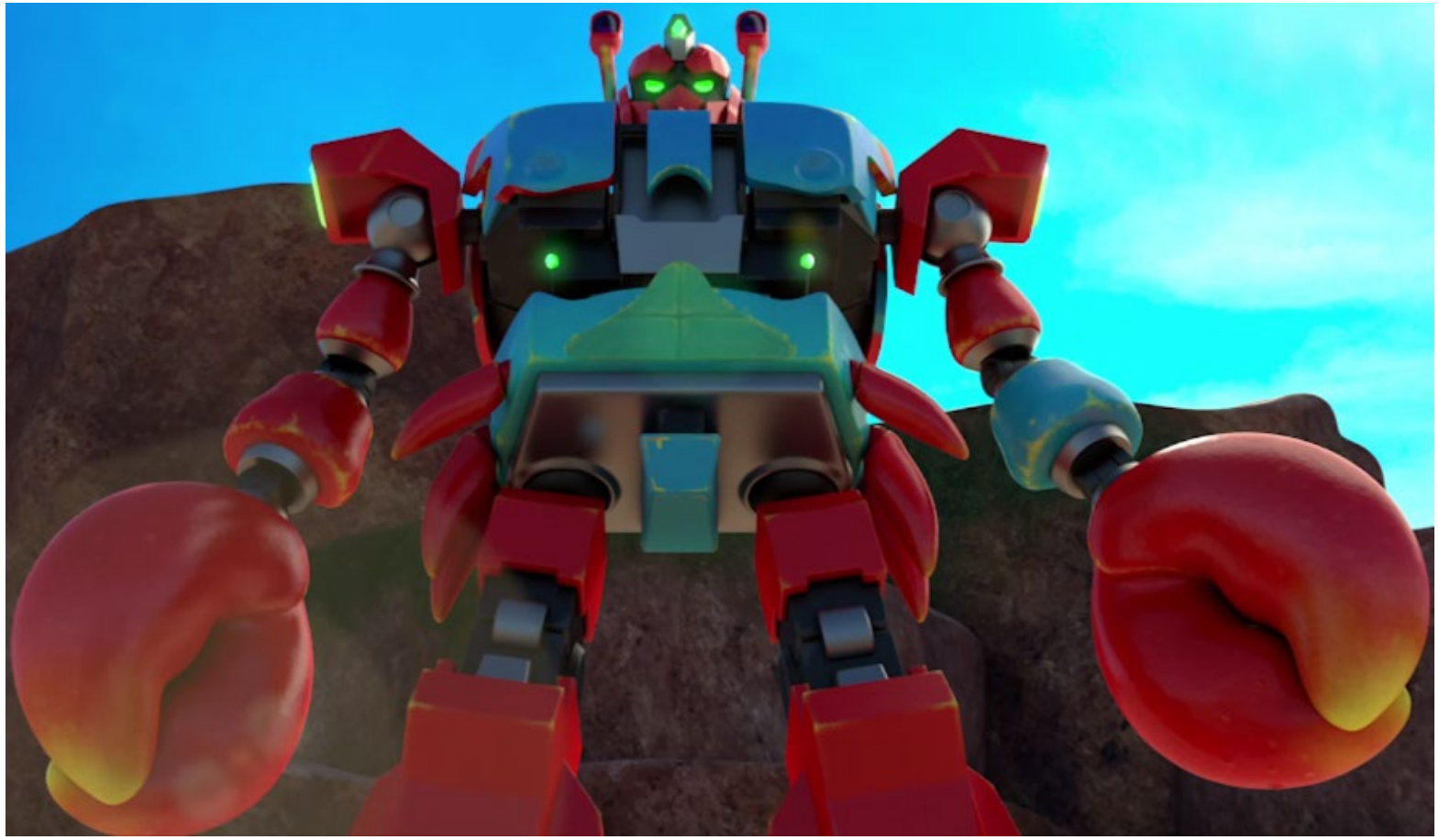
[1]Fachgebiet Keramische Werkstoffe/Chair of Advanced Ceramic Materials, Technische Universität Berlin, Berlin, Germany. [2]Department of Materials Science & Engineering, University of Utah, Salt Lake City, UT, USA. [3]These authors contributed equally: Anthony Yu-Tung Wang, Steven K. Kauwe. ✉email: sparks@eng.utah.edu

Published in partnership with the Shanghai Institute of Ceramics of the Chinese Academy of Sciences

npj nature partner journals

Wang, Kauwe, Murdock, Sparks, npj Comp Mat, 7:77 (2021)

CrabNets are transformers in disguise!



The transformer is behind the NLP revolution

Deepmind's AlphaStar AI

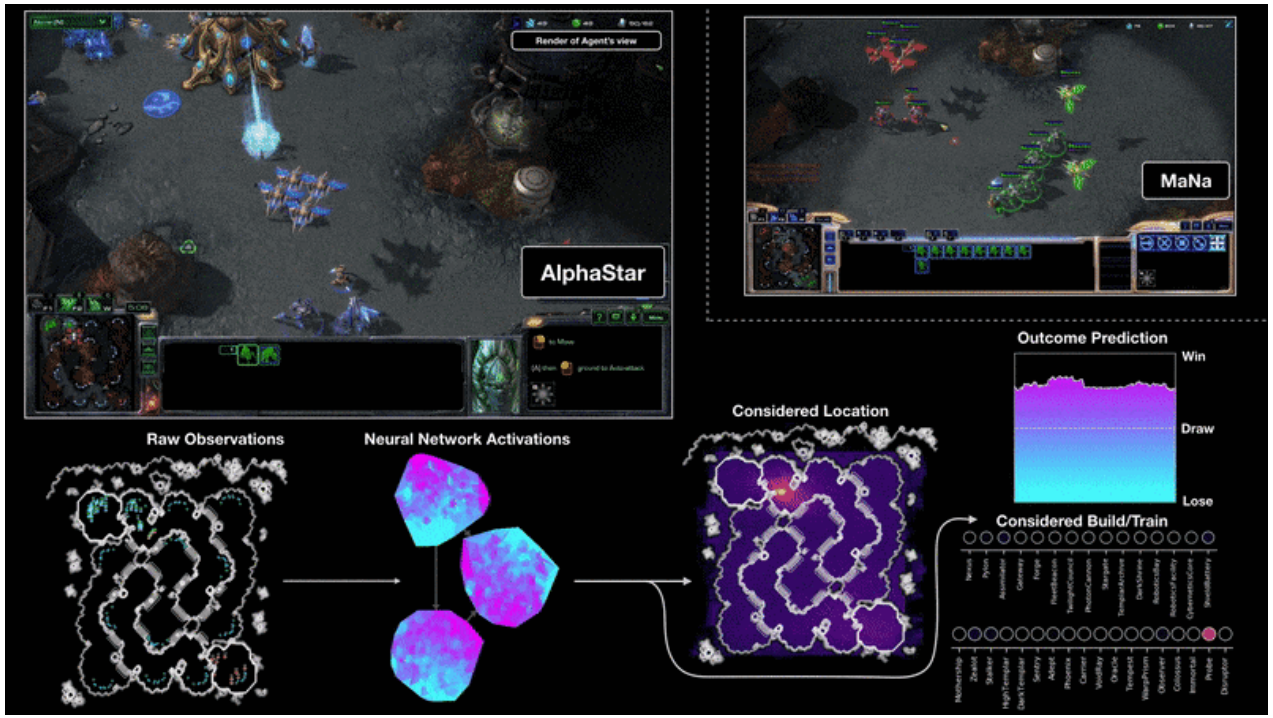


Image GPT (complete the image)

What are the benefits of attention

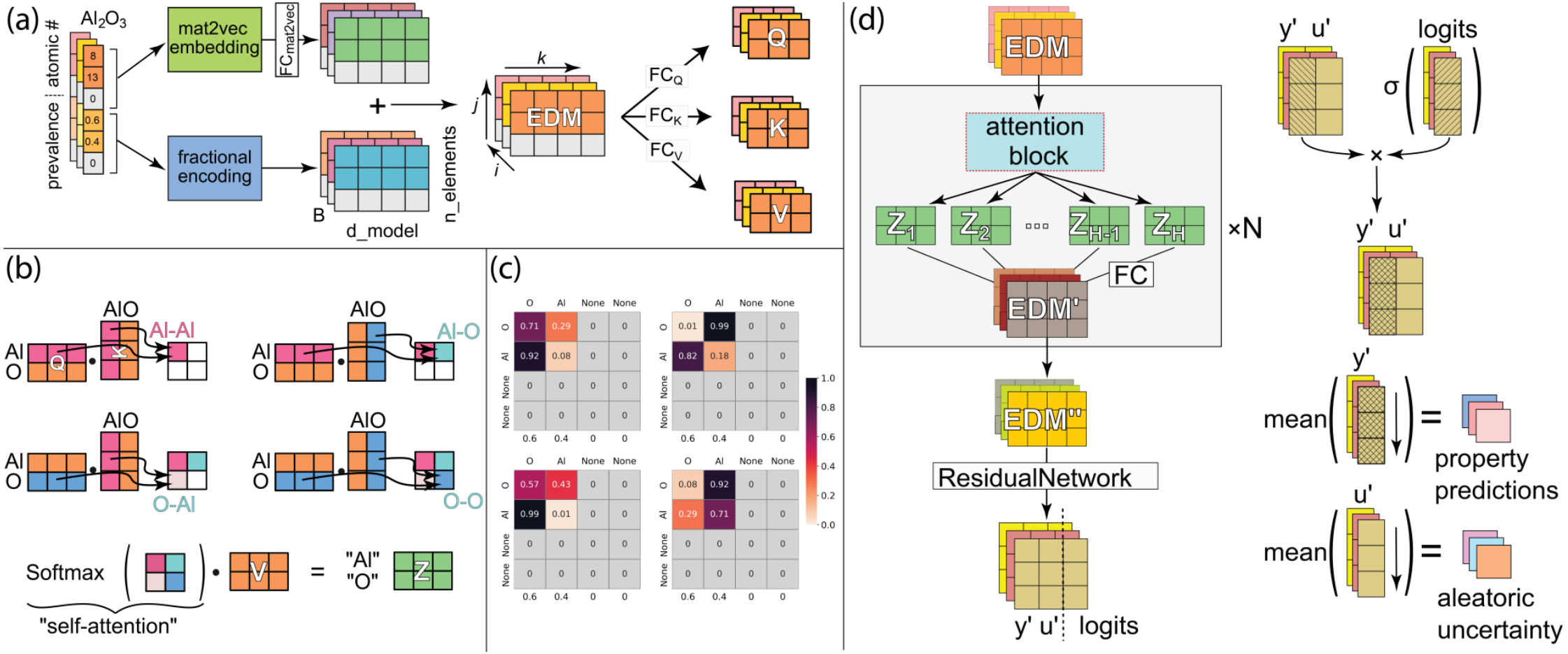
Well optimized learning

Per-element representations

- Element-element interactions
- Attention maps!
- Element property contributions

Permutation invariant

The attention blocks are incorporated into a neural network architecture



Wang, Kauwe, Murdock, Sparks, npj Comp Mat, 7:77 (2021)

CrabNet reaches state-of-the-art performance

Performance comparison

MatBench Properties	Roost	CrabNet	HotCrab	ElemNet	RF
Castelli perovskites	0.359	0.408	0.410	0.468	0.581
Refractive index	0.327	0.309	0.323	0.538	0.419
Shear modulus (log10)	0.104	0.097	0.102	0.140	0.105
Bulk modulus (log10)	0.078	0.073	0.077	0.124	0.082
Experimental band gap	0.374	0.339	0.353	0.450	0.444
DFT Exfoliation energy	47.692	45.617	48.943	57.673	50.000
MP Formation energy	0.082	0.083	0.085	1.029	0.121
MP Band gap	0.252	0.258	0.265	0.337	0.328
Phonon peak	46.158	49.868	56.258	nan	64.924
Steels yield	155.231	91.783	92.507	nan	104.538

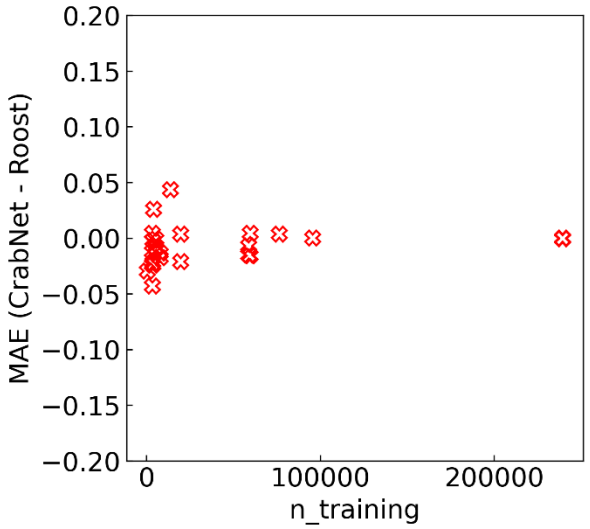
Extended Properties	Roost	CrabNet	HotCrab	ElemNet	RF
AFLOW Bulk modulus	8.988	8.861	9.271	12.242	12.006
AFLOW Debye temperature	37.478	33.717	36.027	47.937	36.352
AFLOW Shear modulus	10.063	9.101	9.510	12.890	10.125
AFLOW Thermal conductivity	2.697	2.316	2.251	3.383	2.724
AFLOW Thermal expansion	3.68e-06	3.84e-06	3.87e-06	6.36e-06	5.54e-06
AFLOW Band gap	0.337	0.302	0.316	0.375	0.384
AFLOW Energy per atom	0.086	0.093	0.094	0.129	0.230
Bartel Decomposition (Ed)	0.067	0.063	0.066	0.081	0.090
Bartel Formation (Ef)	0.055	0.059	0.059	0.071	0.100
MP Bulk modulus	14.950	14.264	14.982	17.236	16.754
MP Elastic anisotropy	9.273	9.233	9.533	9.500	11.239
MP Energy above convex hull	0.089	0.086	0.089	0.100	0.113
MP Magnetic moment	2.583	2.395	2.433	2.761	2.664
MP Shear modulus	12.674	12.363	13.068	15.374	13.368
OQMD Band gap	nan	0.039	0.038	0.119	0.049
OQMD Energy per atom	0.035	0.035	0.035	0.065	0.145
OQMD Formation enthalpy	0.033	0.033	0.032	0.062	0.085
OQMD Volume per atom	0.265	0.247	0.247	nan	0.529

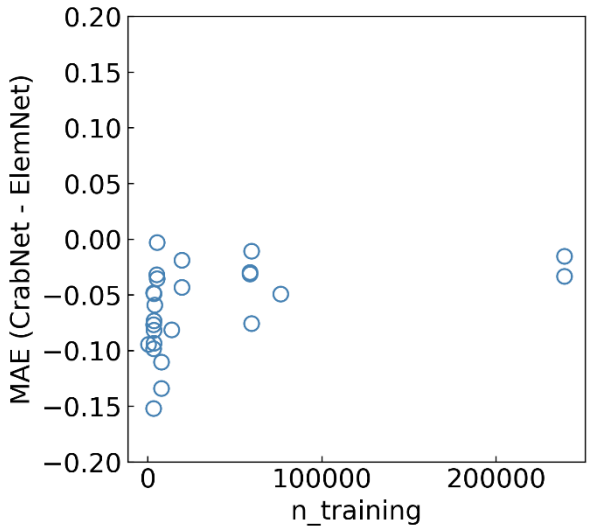
Wang, Kauwe, Murdock, Sparks, npj Comp Mat, 7:77 (2021)

CrabNet reaches state-of-the-art performance

Performance comparison

MatBench Properties	Roost	CrabNet	HotCrab	ElemNet	RF
Castelli perovskites	0.359	0.408	0.410	0.468	0.581
Refractive index	0.327	0.309	0.323	0.538	0.419
Shear modulus (log10)	0.104	0.097	0.102	0.140	0.105
Bulk modulus (log10)	0.078	0.073	0.077	0.124	0.082
Experimental band gap	0.374	0.339	0.353	0.450	0.444
DFT Exfoliation energy	47.692	45.617	48.943	57.673	50.000
MP Formation energy	0.082	0.083	0.085	1.029	0.121
MP Band gap	0.252	0.258	0.265	0.337	0.328
Phonon peak	46.158	49.868	56.258	nan	64.924
Steels yield	155.231	91.783	92.507	nan	104.538

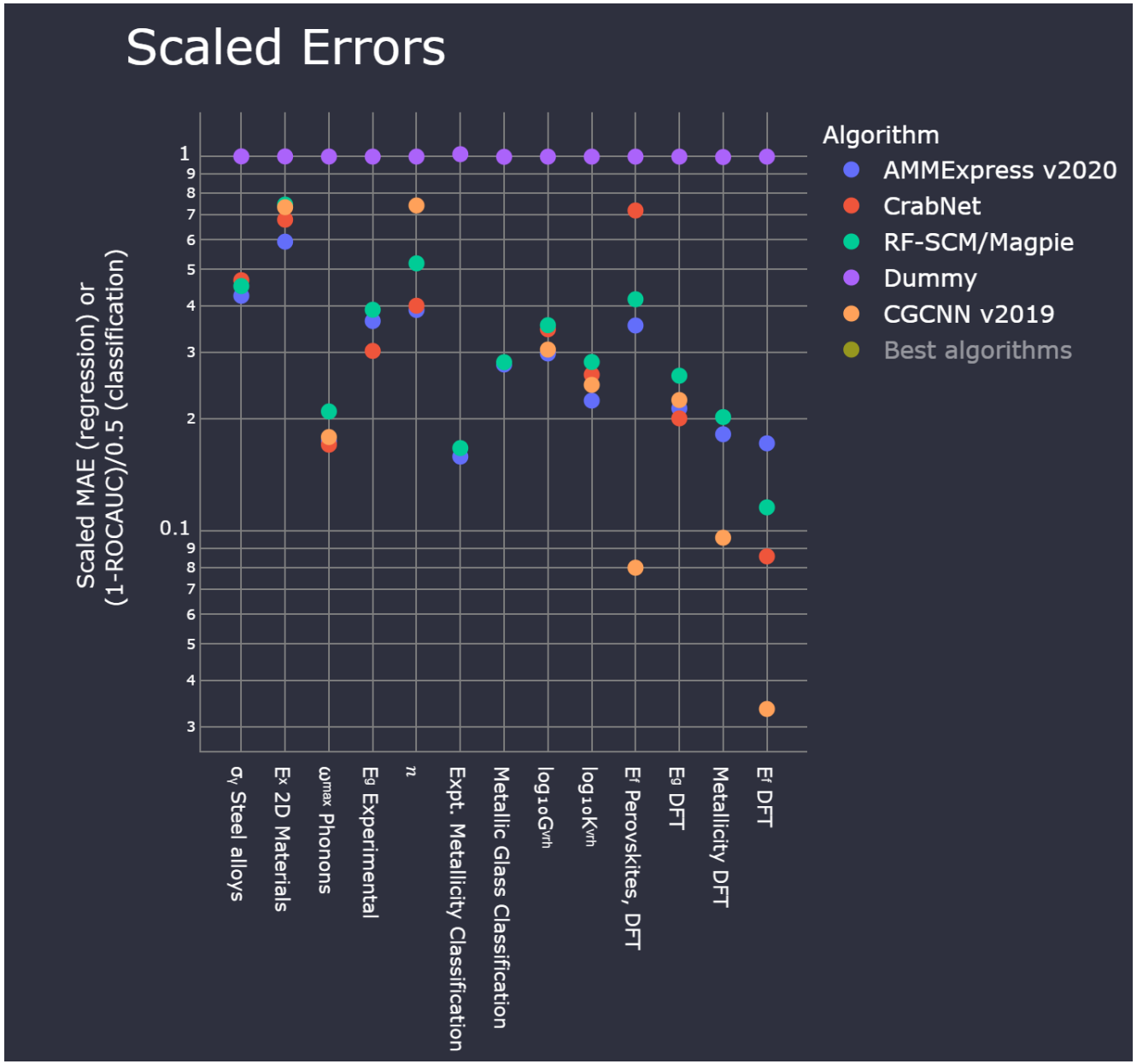




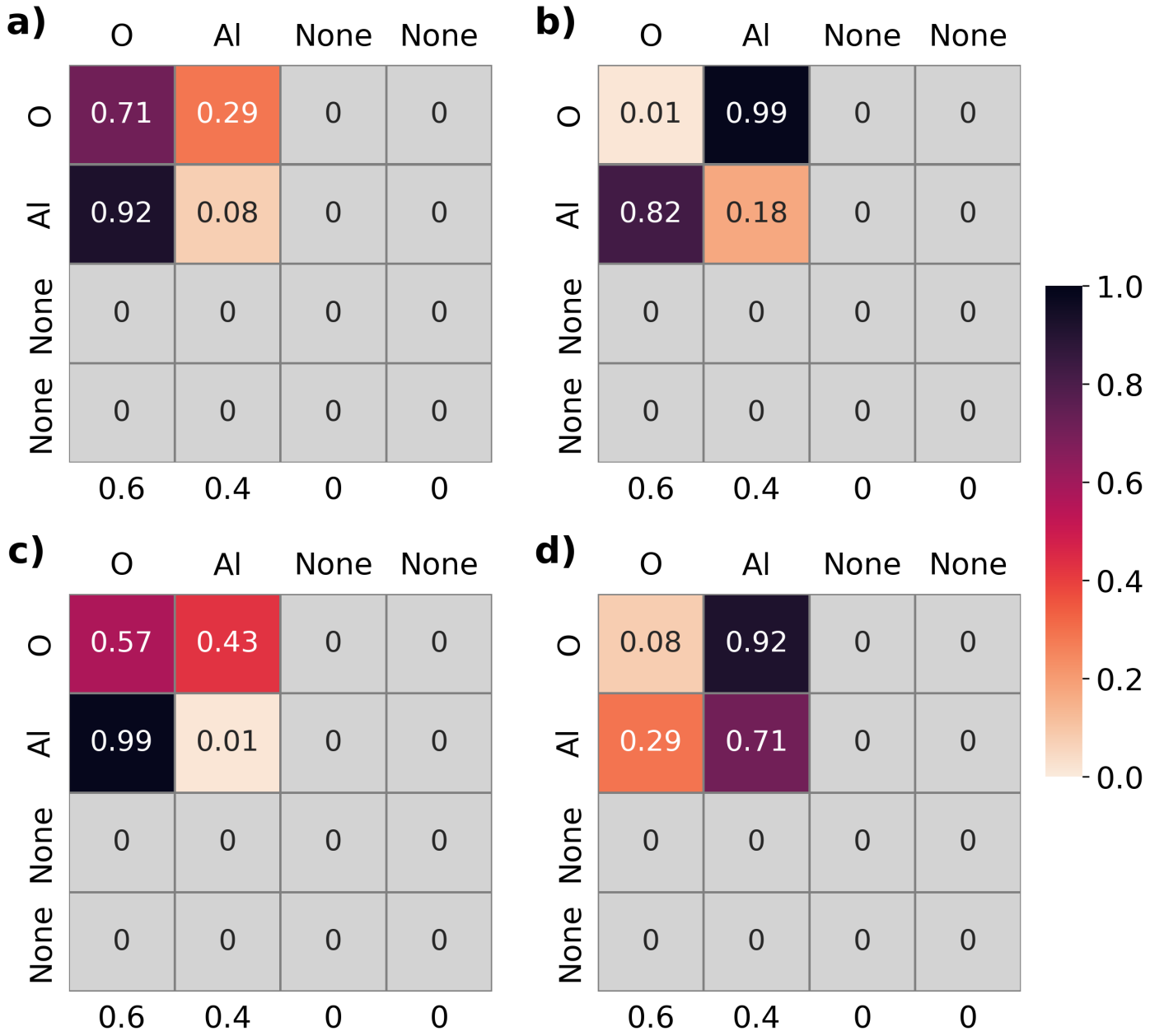
Extended Properties	Roost	CrabNet	HotCrab	ElemNet	RF
AFLOW Bulk modulus	8.988	8.861	9.271	12.242	12.006
AFLOW Debye temperature	37.478	33.717	36.027	47.937	36.352
AFLOW Shear modulus	10.063	9.101	9.510	12.890	10.125
AFLOW Thermal conductivity	2.697	2.316	2.251	3.383	2.724
AFLOW Thermal expansion	3.68e-06	3.84e-06	3.87e-06	6.36e-06	5.54e-06
AFLOW Band gap	0.337	0.302	0.316	0.375	0.384
AFLOW Energy per atom	0.086	0.093	0.094	0.129	0.230
Bartel Decomposition (Ed)	0.067	0.063	0.066	0.081	0.090
Bartel Formation (Ef)	0.055	0.059	0.059	0.071	0.100
MP Bulk modulus	14.950	14.264	14.982	17.236	16.754
MP Elastic anisotropy	9.273	9.233	9.533	9.500	11.239
MP Energy above convex hull	0.089	0.086	0.089	0.100	0.113
MP Magnetic moment	2.583	2.395	2.433	2.761	2.664
MP Shear modulus	12.674	12.363	13.068	15.374	13.368
OQMD Band gap	nan	0.039	0.038	0.119	0.049
OQMD Energy per atom	0.035	0.035	0.035	0.065	0.145
OQMD Formation enthalpy	0.033	0.033	0.032	0.062	0.085
OQMD Volume per atom	0.265	0.247	0.247	nan	0.529

Wang, Kauwe, Murdock, Sparks, npj Comp Mat, 7:77 (2021)

CrabNet on par with algorithms that include both structure and composition!



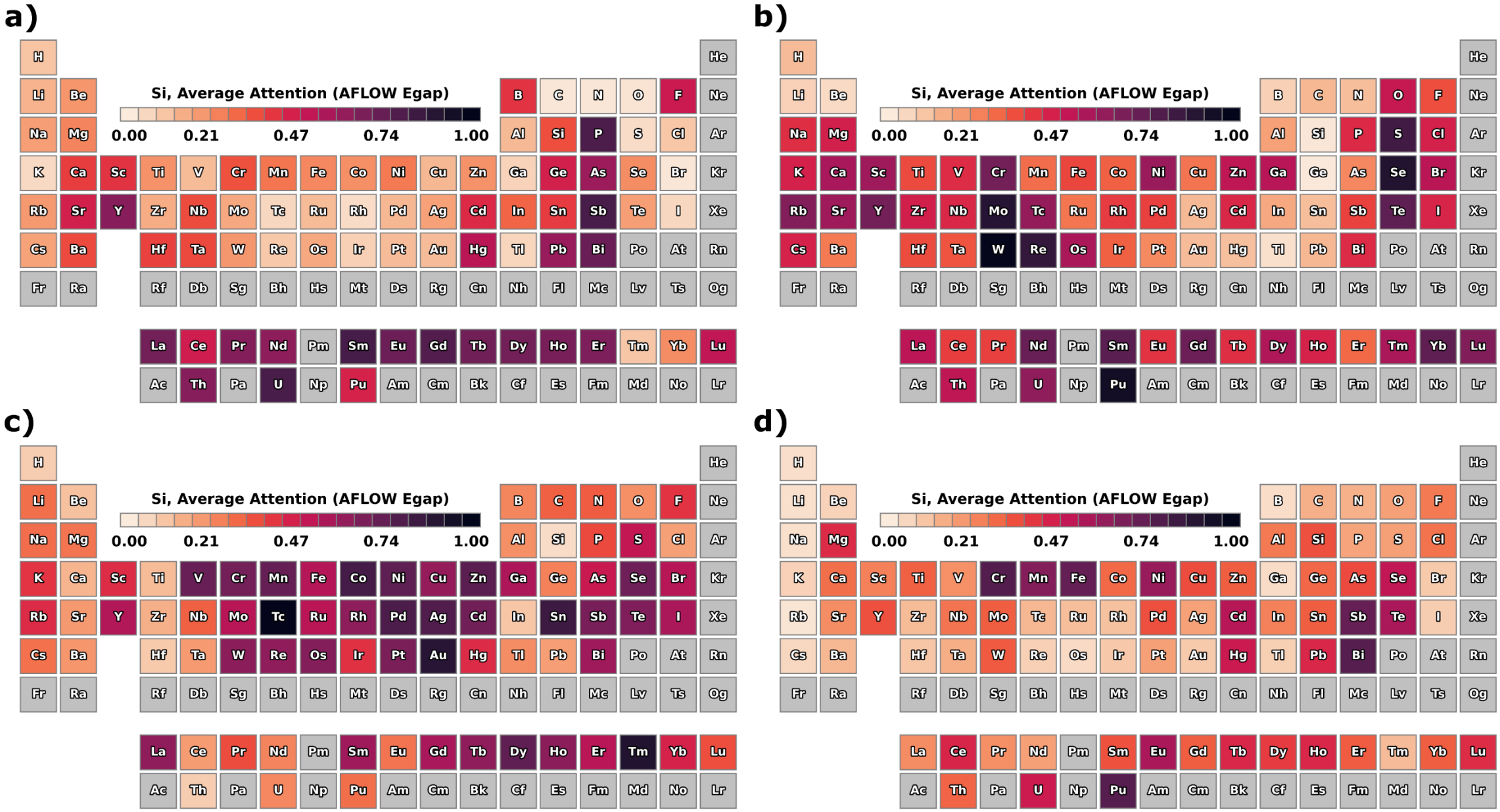
Attention maps allow us to visualize where the algorithms attends to for predictions



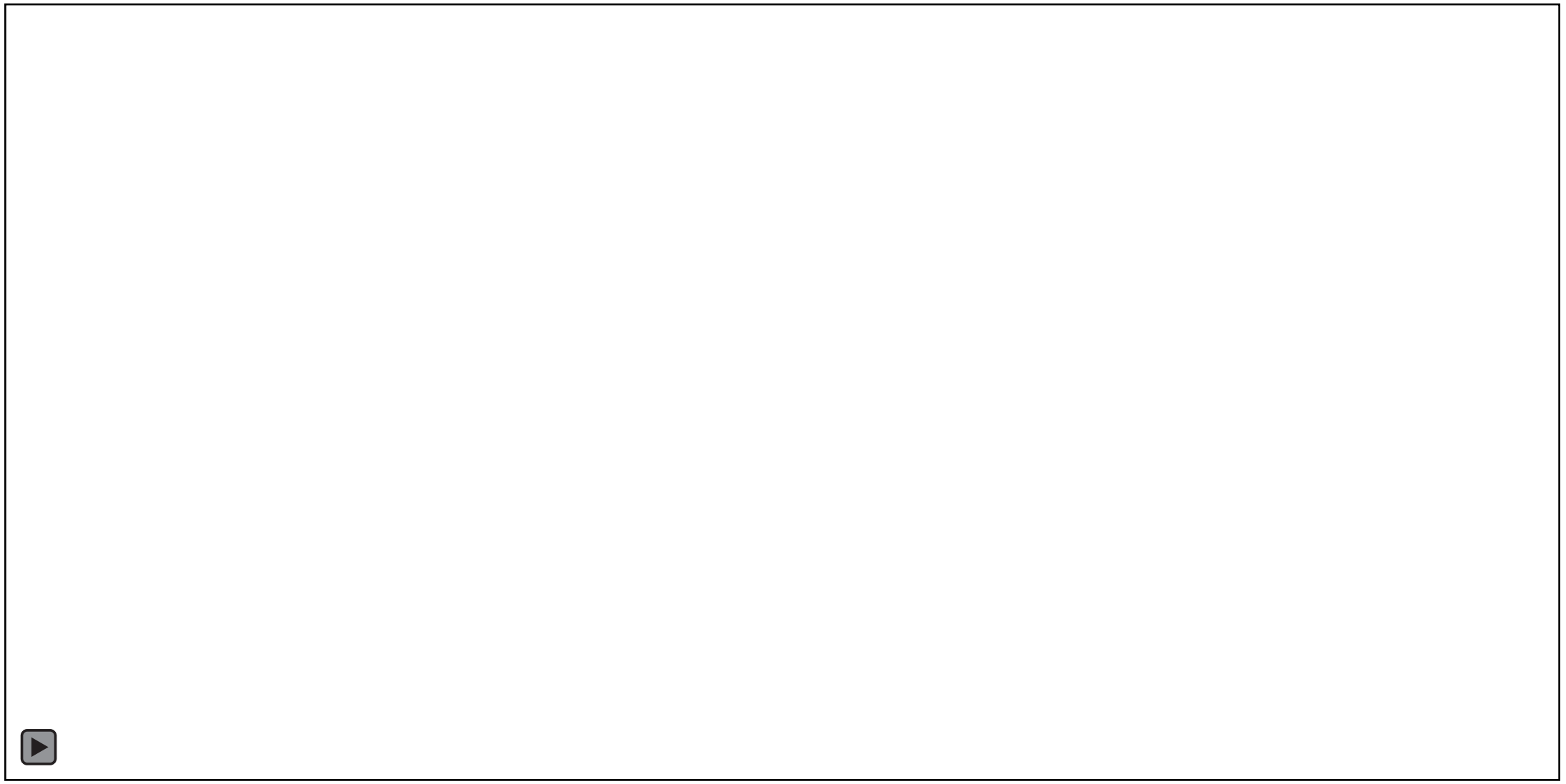
Al_2O_3 Bulk Modulus

Attention heads ***after training***.

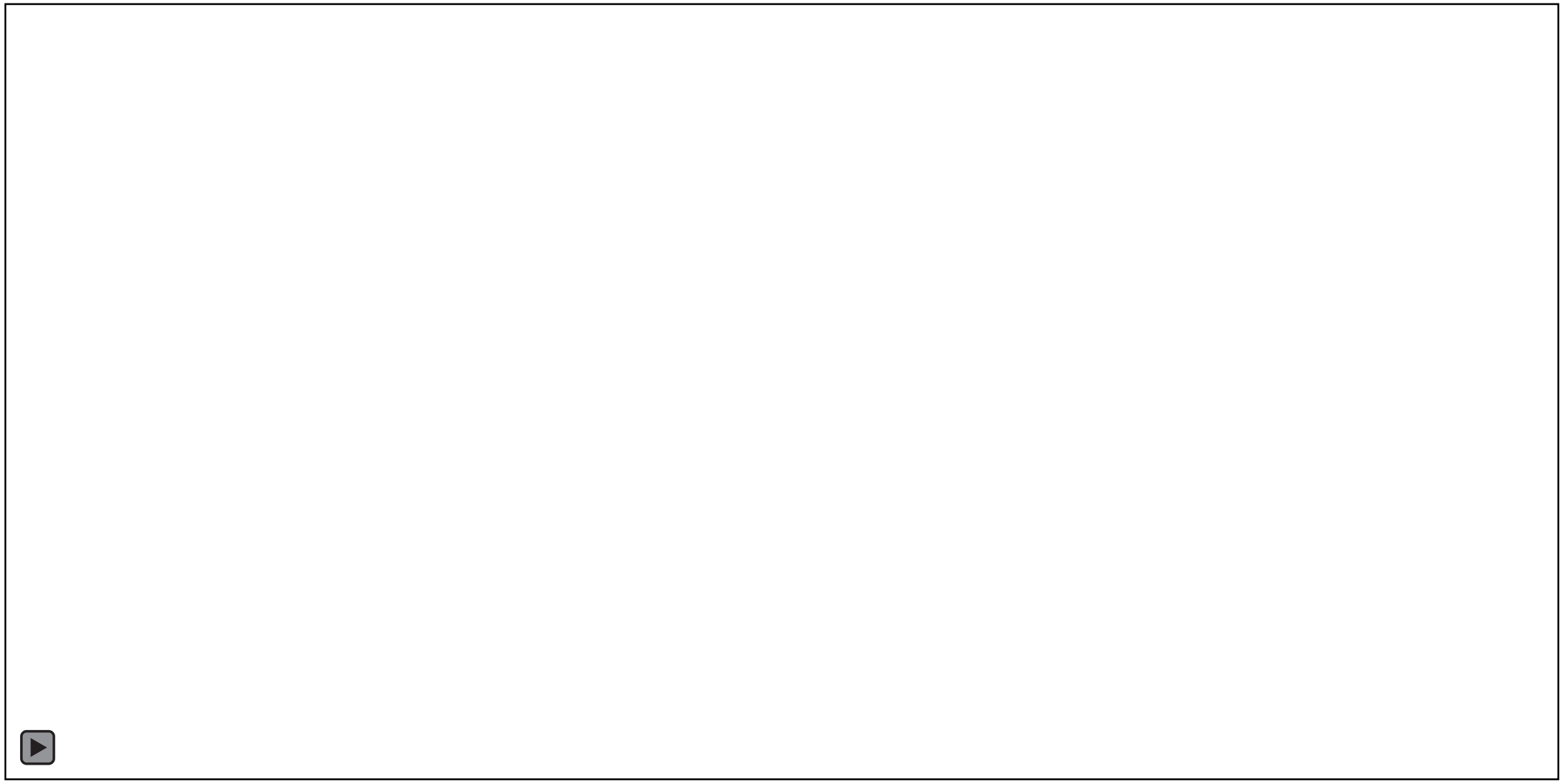
Attention maps allow us to visualize where the algorithms attends to for predictions



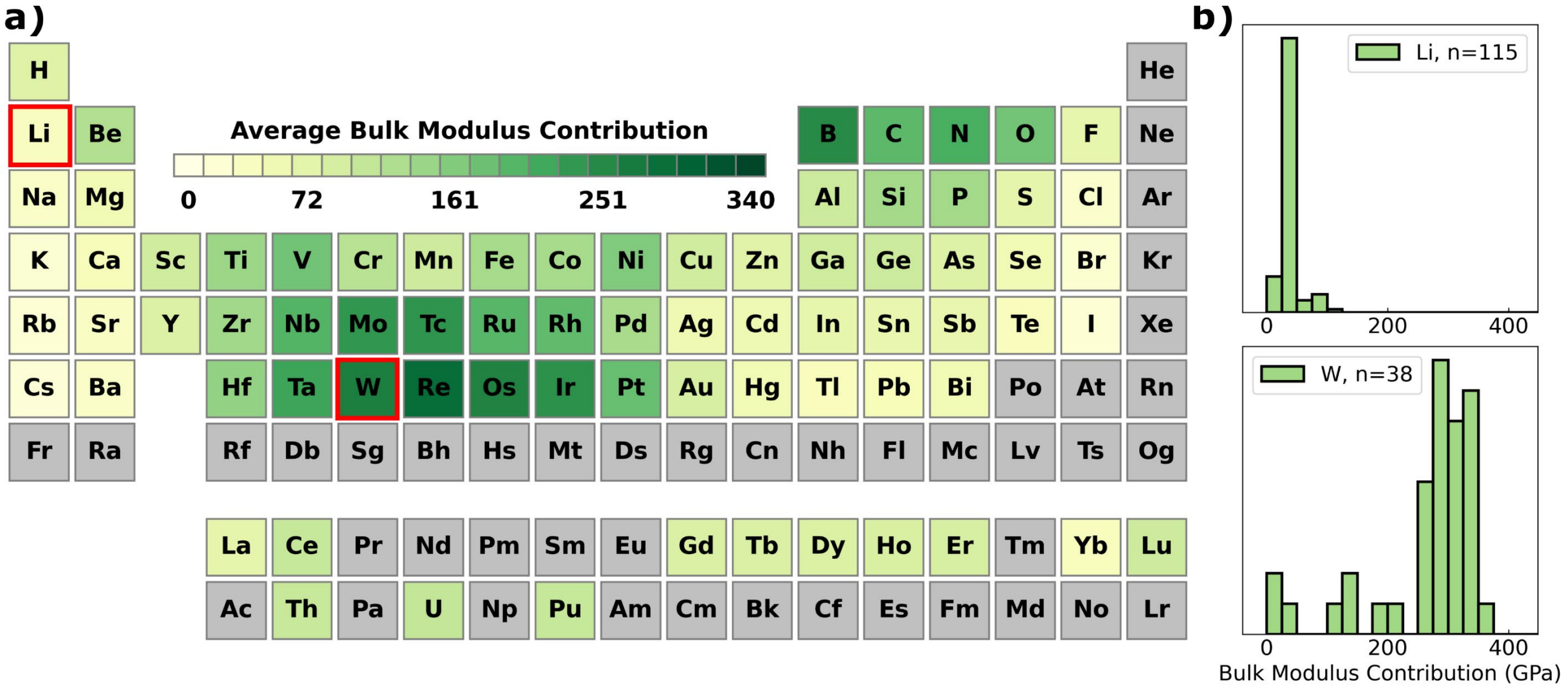
Maps are generated during training, so we can create videos of how they change



Maps are generated during training, so we can create videos of how they change

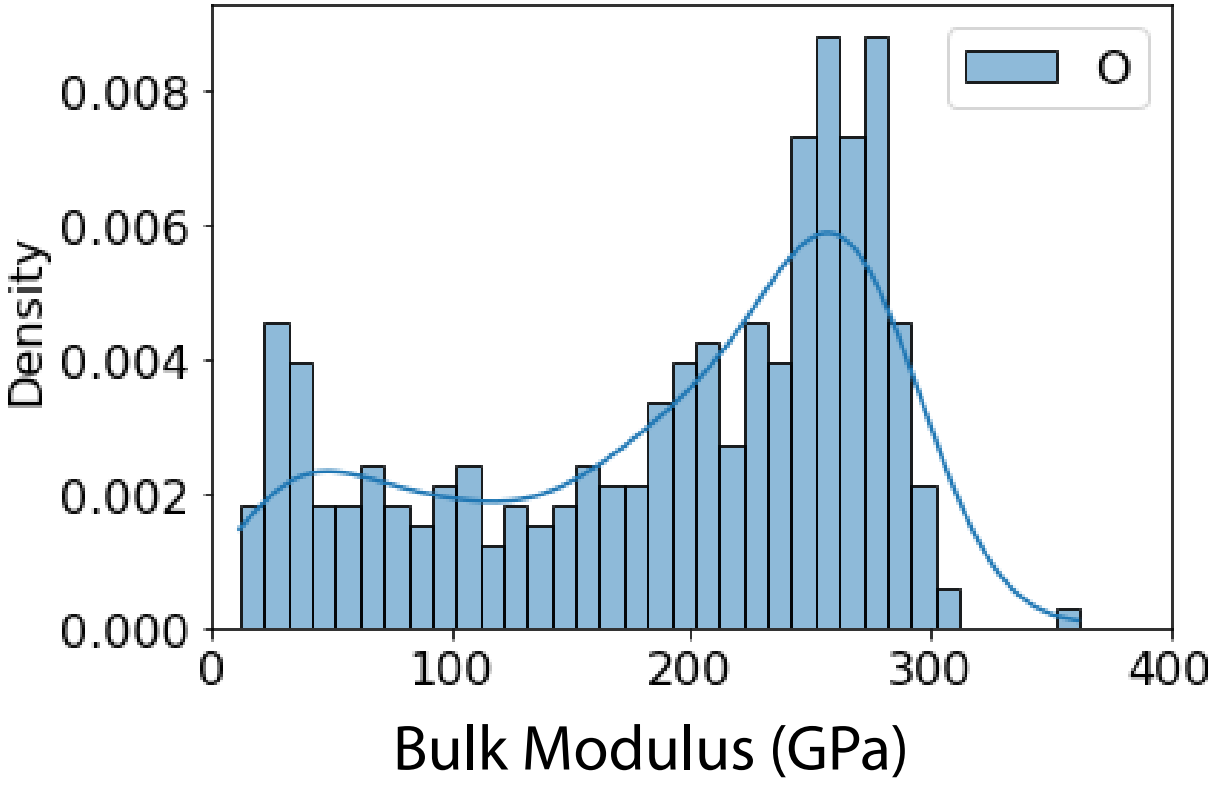


Visualization of element contributions offers better understanding and better expectations

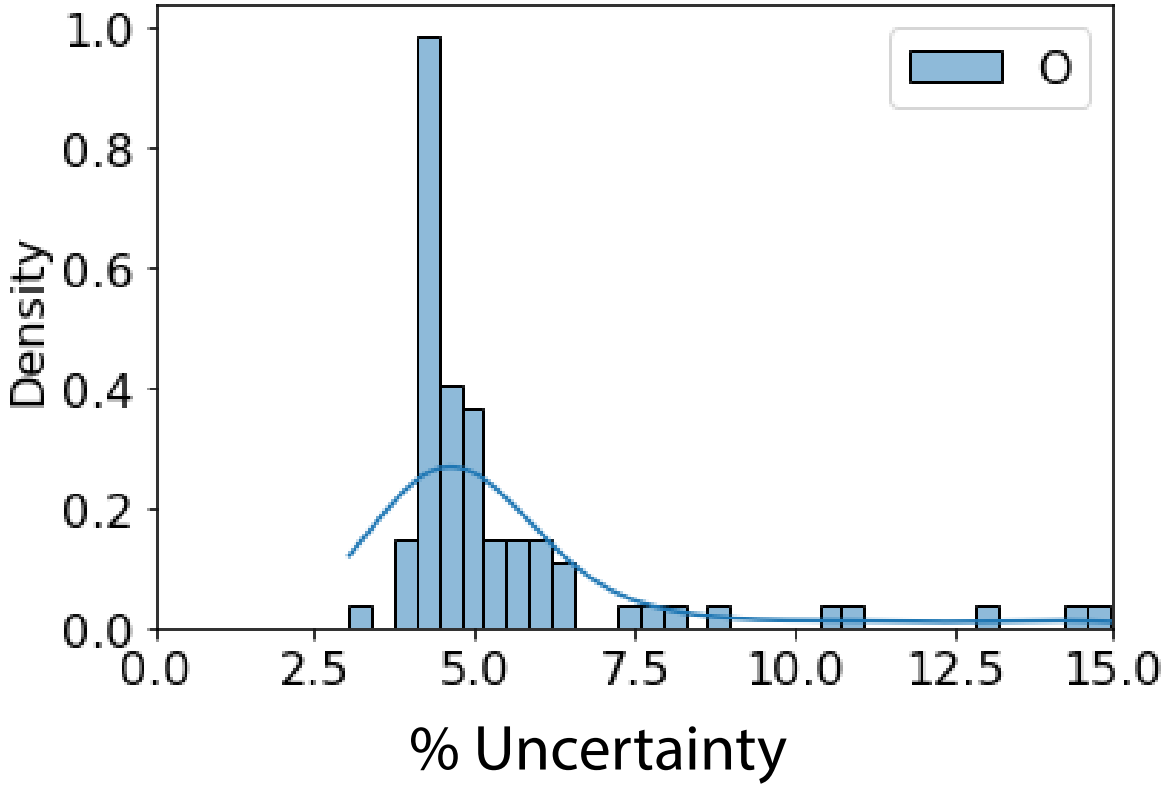


We can also see elemental contribution to a property and uncertainty

Element bulk modulus contribution



Element uncertainty contribution



generative machine learning models