AstraZeneca

Using test-time augmentation to investigate XAI: inconsistencies between method, model and human intuition

P.B.R. Hartog, F. Krüger, S. Genheden, I.V. Tetko

22 april 2024

We've all seen it before

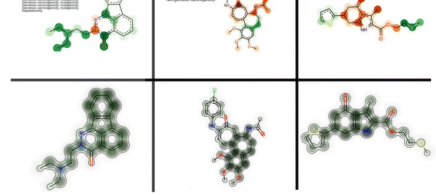


Fig. 6 Cytotoxicity case studies. Three molecules from the FMP dataset are visualized, using either their attention scores from our method (top) or their cytotoxicity maps using deep Taylor decomposition (bottom) following the original work.[30] For our CLM, the color mapping is identical to Fig. 3: green for toxicophore atoms and orange for non-toxicophore atoms. Opacity corresponds to the attention score.

- A new state-of-the-art model

Table 2 ROC-AUC values on the MoleculeNet datasets for different algorithms. With the exception of ClinTox, our method always obtained either the best (bold) or second-best (italic) performance on each dataset. Across all datasets, we outperform all competing approaches. For each dataset, ten models were trained and the splitting strategy recommended by MoleculeNet was utilized

Dataset	**BACE**	**SIDER**	**Clintox**	**BBBP**	**Tox21**	**HIV**	
Split	**Scaffold**	**Random**	**Random**	**Scaffold**	**Random**	**Scaffold**	**Average**
Ours (CLM)	**0.861**$_{\pm 0.04}$	*0.659*$_{\pm 0.04}$	0.878$_{\pm 0.00}$	**0.915**$_{\pm 0.02}$	**0.858**$_{\pm 0.05}$	**0.813**$_{\pm 0.03}$	**0.831**
GraphConv (Wu Z. *et al.*)[11]	0.783$_{\pm 0.01}$	0.638$_{\pm 0.01}$	0.807$_{\pm 0.05}$	0.690$_{\pm 0.01}$	0.829$_{\pm 0.01}$	0.763$_{\pm 0.02}$	0.752
Weave (Wu Z. *et al.*)[11]	0.806$_{\pm 0.00}$	0.581$_{\pm 0.03}$	0.832$_{\pm 0.04}$	0.671$_{\pm 0.01}$	0.820$_{\pm 0.01}$	0.703$_{\pm 0.04}$	0.736
D-MPNN (Yang K. *et al.*)[12]	0.838$_{\pm 0.06}$	0.646$_{\pm 0.02}$	**0.894**$_{\pm 0.03}$	*0.888*±0.03	*0.845*$_{\pm 0.002}$	*0.794*$_{\pm 0.02}$	*0.818*
SELU-MPNN (Withnall M. *et al.*)[103]	—	0.632$_{\pm 0.01}$	—	0.693$_{\pm 0.06}$	0.820$_{\pm 0.01}$	0.747$_{\pm 0.01}$	—
AMPNN (Withnall M. *et al.*)[103]	—	0.639$_{\pm 0.01}$	—	0.709$_{\pm 0.04}$	0.812$_{\pm 0.02}$	0.742$_{\pm 0.02}$	—
EMPNN (Withnall M. *et al.*)[103]	—	0.651$_{\pm 0.01}$	—	0.705$_{\pm 0.02}$	0.829$_{\pm 0.01}$	0.759$_{\pm 0.01}$	—
MMNB (Shen W. X. *et al.*)[104]	*0.849*	**0.680**	*0.888*	0.739	0.842	0.777	0.796

(1) Jannis Born, Greta Markert, Nikita Janakarajan, Talia B Kimber, Andrea Volkamer, María Rodríguez Martínez, and Matteo Manica. Chemical representation learning for toxicity prediction. Digital Discovery, 2023.

We've all seen it before

- A new state-of-the-art model
- Publication uses XAI

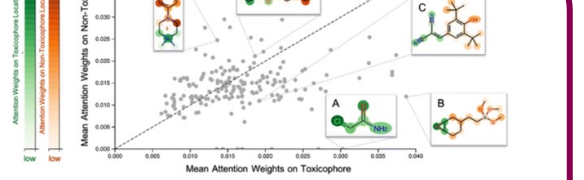


Fig. 3 Visualization of attention maps of Tox21 compounds. Scatterplot of the mean attention of the toxicophore part of the molecule *versus* the mean attention of the remaining ones. Toxicophoric atoms are colored green, the remaining parts are colored red, and intensity encodes the attention weight. Toxicophores are assigned significantly higher attention weights. Interactive visualization is available at: https://ibm.biz/tox21_attention. MWU denotes a Mann–Whitney *U* test.

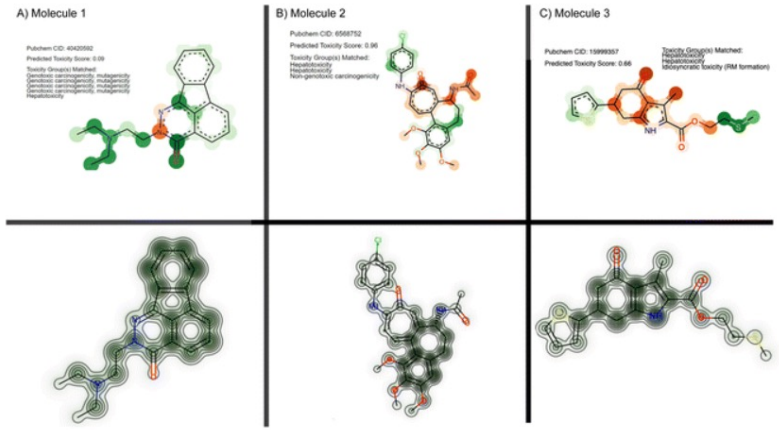


Fig. 6 Cytotoxicity case studies. Three molecules from the FMP dataset are visualized, using either their attention scores from our method (top) or their cytotoxicity maps using deep Taylor decomposition (bottom) following the original work.[30] For our CLM, the color mapping is identical to Fig. 3: green for toxicophore atoms and orange for non-toxicophore atoms. Opacity corresponds to the attention score.

Table 2 ROC-AUC values on the MoleculeNet datasets for different algorithms. With the exception of ClinTox, our method always obtained either the best (bold) or second-best (italic) performance on each dataset. Across all datasets, we outperform all competing approaches. For each dataset, ten models were trained and the splitting strategy recommended by MoleculeNet was utilized

Dataset	BACE	SIDER	Clintox	BBBP	Tox21	HIV	
Split	Scaffold	Random	Random	Scaffold	Random	Scaffold	Average
Ours (CLM)	**0.861**$_{\pm 0.04}$	*0.659*$_{\pm 0.04}$	0.878$_{\pm 0.00}$	**0.915**$_{\pm 0.02}$	**0.858**$_{\pm 0.05}$	**0.813**$_{\pm 0.03}$	**0.831**
GraphConv (Wu Z. *et al.*)[11]	0.783$_{\pm 0.01}$	0.638$_{\pm 0.01}$	0.807$_{\pm 0.05}$	0.690$_{\pm 0.01}$	0.829$_{\pm 0.01}$	0.763$_{\pm 0.02}$	0.752
Weave (Wu Z. *et al.*)[11]	0.806$_{\pm 0.00}$	0.581$_{\pm 0.03}$	0.832$_{\pm 0.04}$	0.671$_{\pm 0.01}$	0.820$_{\pm 0.01}$	0.703$_{\pm 0.04}$	0.736
D-MPNN (Yang K. *et al.*)[12]	0.838$_{\pm 0.06}$	0.646$_{\pm 0.02}$	**0.894**$_{\pm 0.03}$	*0.888*±0.03	*0.845*$_{\pm 0.002}$	*0.794*$_{\pm 0.02}$	*0.818*
SELU-MPNN (Withnall M. *et al.*)[103]	—	0.632$_{\pm 0.01}$	—	0.693$_{\pm 0.06}$	0.820$_{\pm 0.01}$	0.747$_{\pm 0.01}$	—
AMPNN (Withnall M. *et al.*)[103]	—	0.639$_{\pm 0.01}$	—	0.709$_{\pm 0.04}$	0.812$_{\pm 0.02}$	0.742$_{\pm 0.02}$	—
EMPNN (Withnall M. *et al.*)[103]	—	0.651$_{\pm 0.01}$	—	0.705$_{\pm 0.02}$	0.829$_{\pm 0.01}$	0.759$_{\pm 0.01}$	—
MMNB (Shen W. X. *et al.*)[104]	*0.849*	**0.680**	*0.888*	0.739	0.842	0.777	0.796

(1) Jannis Born, Greta Markert, Nikita Janakarajan, Talia B Kimber, Andrea Volkamer, María Rodríguez M
representation learning for toxicity prediction. Digital Discovery, 2023.

We've all seen it before

- A new state-of-the-art model
- Publication uses XAI
- It agrees with expert opinion

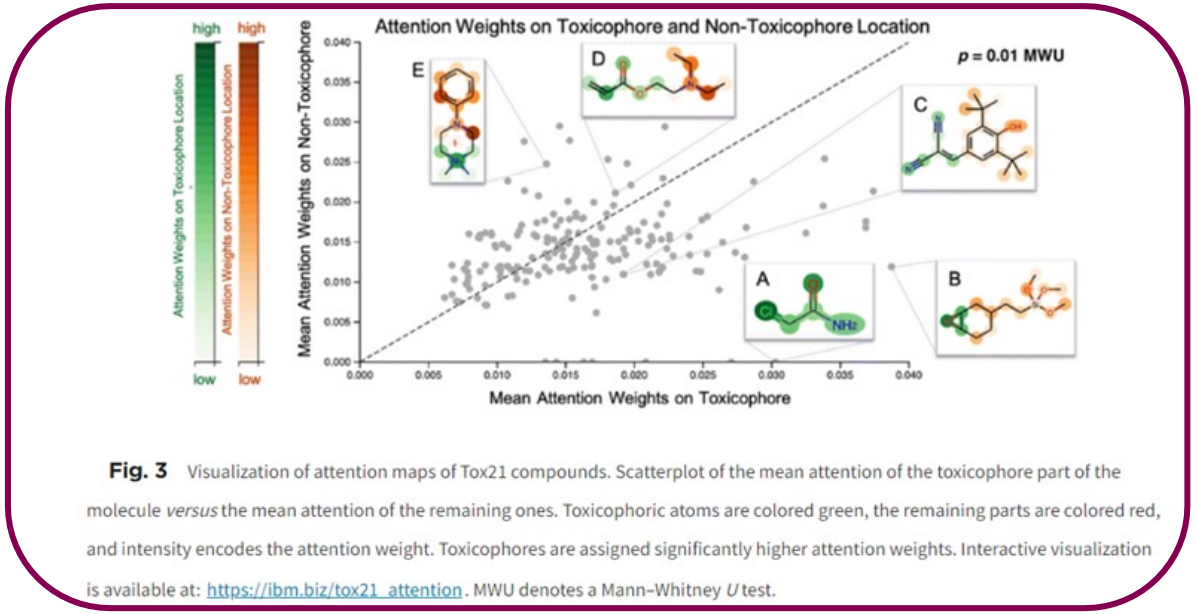


Fig. 3 Visualization of attention maps of Tox21 compounds. Scatterplot of the mean attention of the toxicophore part of the molecule *versus* the mean attention of the remaining ones. Toxicophoric atoms are colored green, the remaining parts are colored red, and intensity encodes the attention weight. Toxicophores are assigned significantly higher attention weights. Interactive visualization is available at: https://ibm.biz/tox21_attention. MWU denotes a Mann–Whitney *U* test.

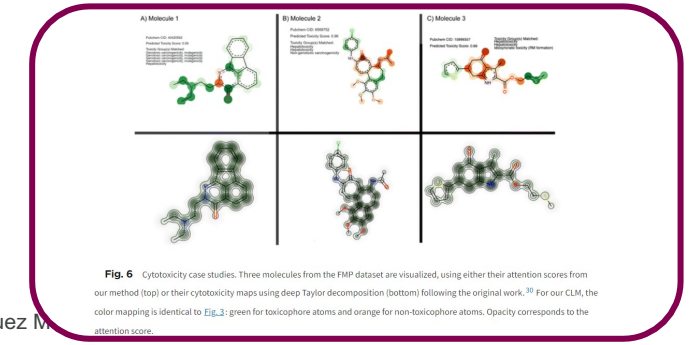


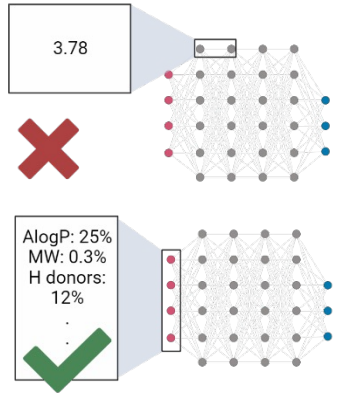
Fig. 6 Cytotoxicity case studies. Three molecules from the FMP dataset are visualized, using either their attention scores from our method (top) or their cytotoxicity maps using deep Taylor decomposition (bottom) following the original work.[30] For our CLM, the color mapping is identical to Fig. 3: green for toxicophore atoms and orange for non-toxicophore atoms. Opacity corresponds to the attention score.

(1) Jannis Born, Greta Markert, Nikita Janakarajan, Talia B Kimber, Andrea Volkamer, María Rodríguez M
representation learning for toxicity prediction. Digital Discovery, 2023.

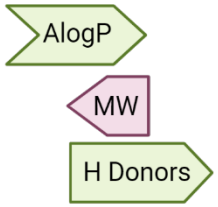
XAI Checklist

Human Interpretable

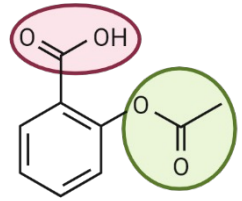
Interpretable



Quickly understood

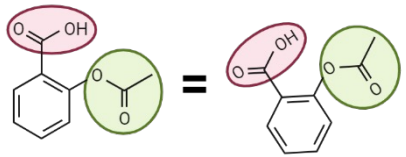


Preferably visual

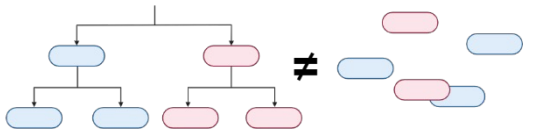


Robust

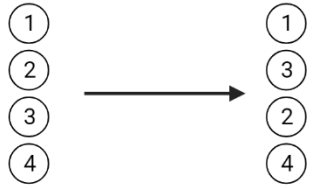
Consistent



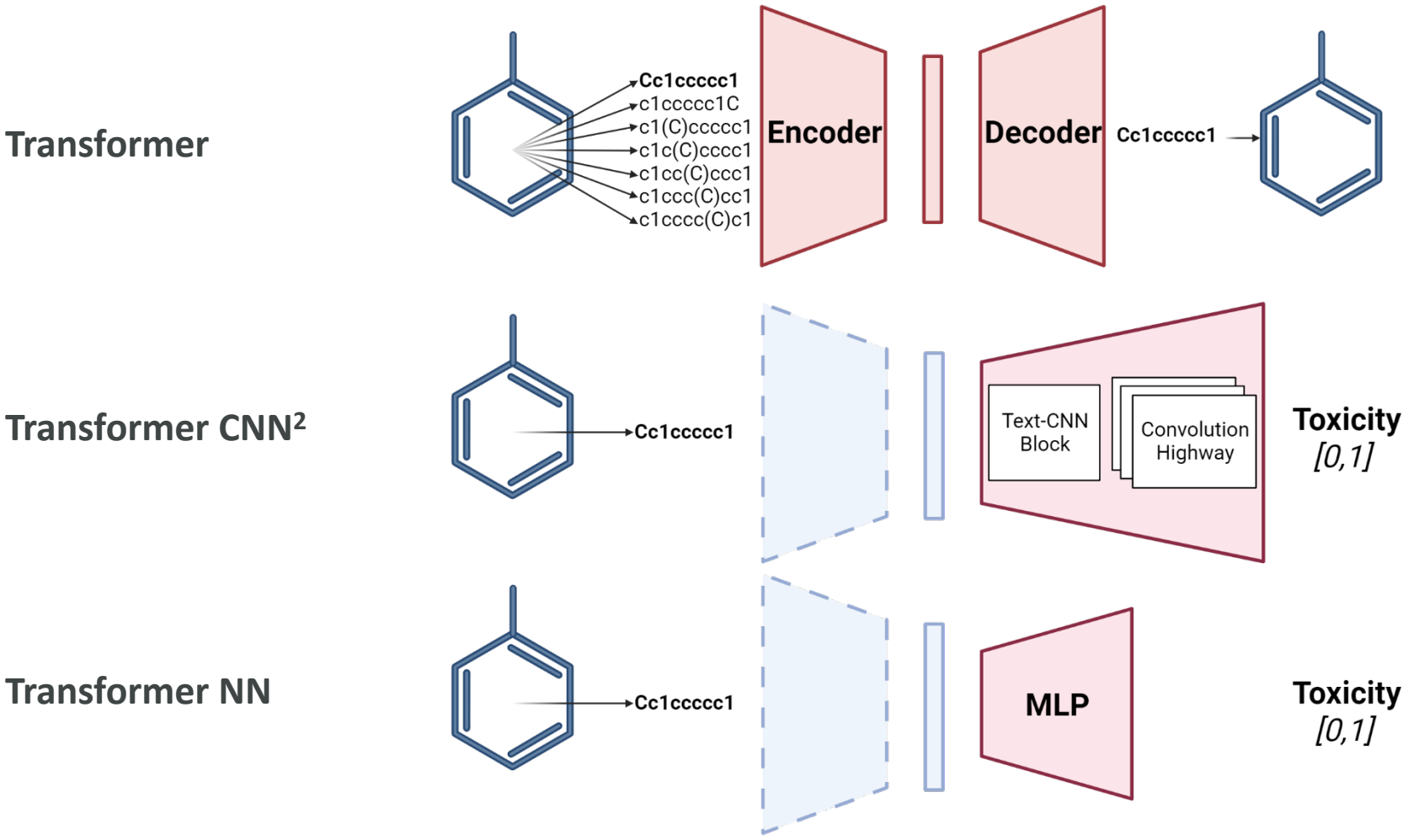
Better than random



Consistent ranking



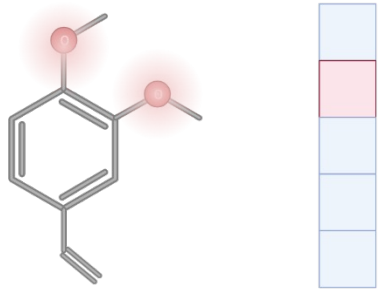
NLP-based Molecular Representation



(2) Pavel Karpov, Guillaume Godin, and Igor V Tetko. Transformer-cnn: Swiss knife for qsar modeling and interpretation. Journal of cheminformatics, 12(1):1–12, 2020.

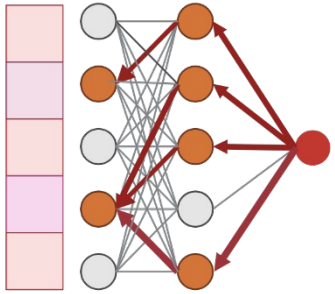
XAI Methods

Pertubation-based XAI



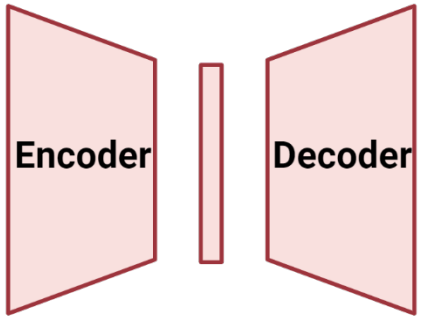
SHAP[3]

Gradient-based XAI



Integrated Gradients (IG)[4]

Transformer-dependent XAI



Attention Maps, Rollout, Grads, AttGrads, CAT and AttCAT[5]

(3) Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. " why should i trust you?" explaining the predictions of any classifier. In Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining, pages 1135–1144, 2016.
(4) Mukund Sundararajan, Ankur Taly, and Qiqi Yan. Axiomatic attribution for deep networks, 2017.
(5) Yao Qiang, Deng Pan, Chengyin Li, Xin Li, Rhongho Jang, and Dongxiao Zhu. Attcat: Explaining trans-formers via attentive class activation tokens. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho,and A. Oh, editors, Advances in Neural Information Processing Systems, volume 35, pages 5052–5064. Cur-ran Associates, Inc., 2022. URL https://proceedings.neurips.cc/paper_files/paper/2022/file/20e45668fefa793bd9f2edf19be12c4b-Paper-Conference.pdf.

Ames Mutagenicity Test

- From Therapeutics Data Commons[6]

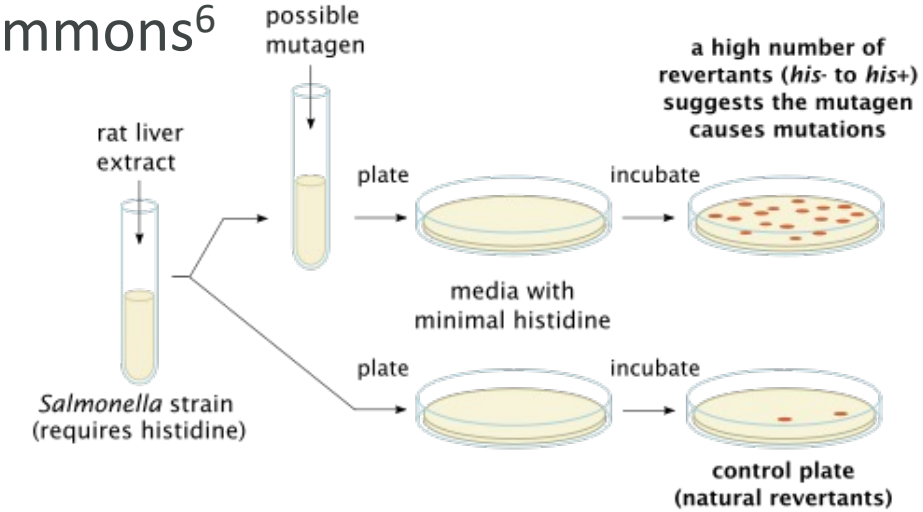


Image from: https://en.wikipedia.org/wiki/Ames_test

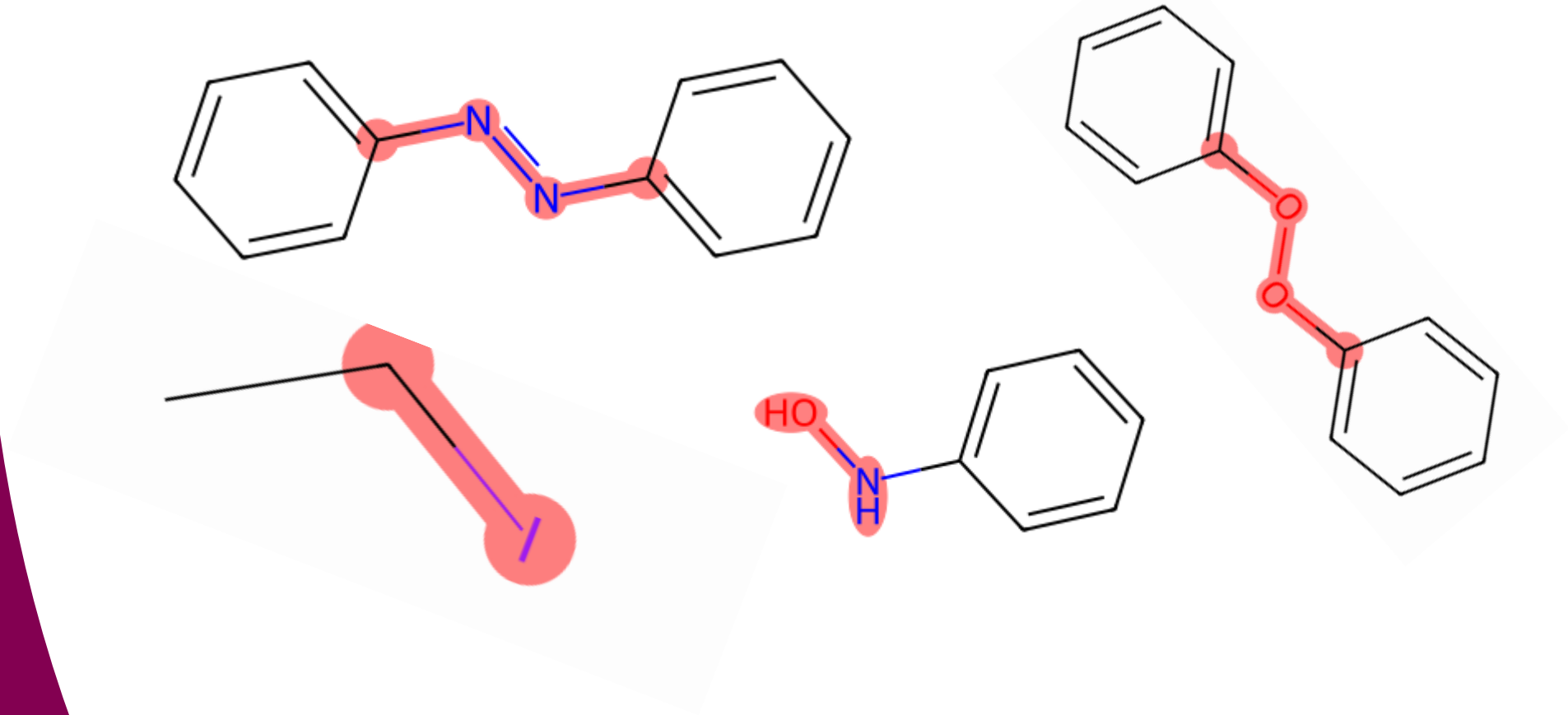
- Predefined scaffold split:

	Training	Validation	Test
Toxic	2470	391	776
Non-toxic	2141	297	647
Total:	4611	688	1423

(6) Kexin Huang, Tianfan Fu, Wenhao Gao, Yue Zhao, Yusuf Roohani, Jure Leskovec, Connor W Coley, Cao Xiao, Jimeng Sun, and Marinka Zitnik. Therapeutics data commons: Machine learning datasets and tasks for drug discovery and development. arXiv preprint arXiv:2102.09548, 2021.

Structural Alerts

- Ames Expert-derived structural alerts[7] (52)



(7) Jeroen Kazius, Ross McGuire, and Roberta Bursi. Derivation and validation of toxicophores for mutagenicity prediction. Journal of medicinal chemistry, 48(1):312–320, 2005.

Enumeration Variance

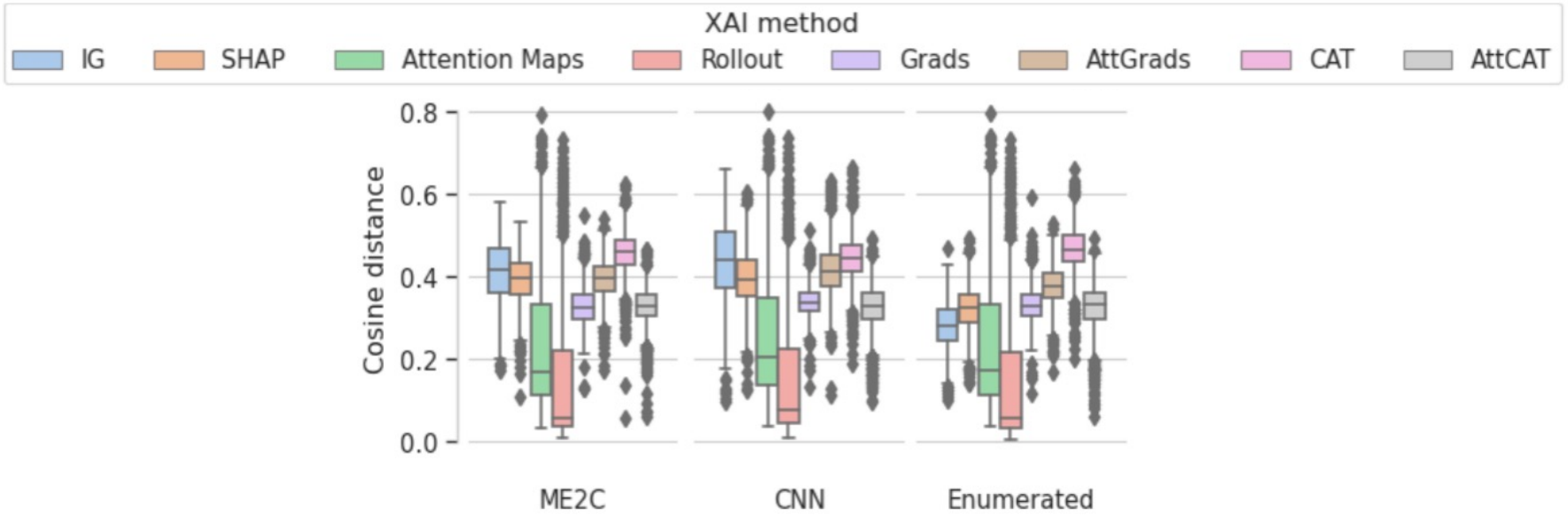


Figure 3: **In-between sample cosine distance of different training settings.**

Information is conserved

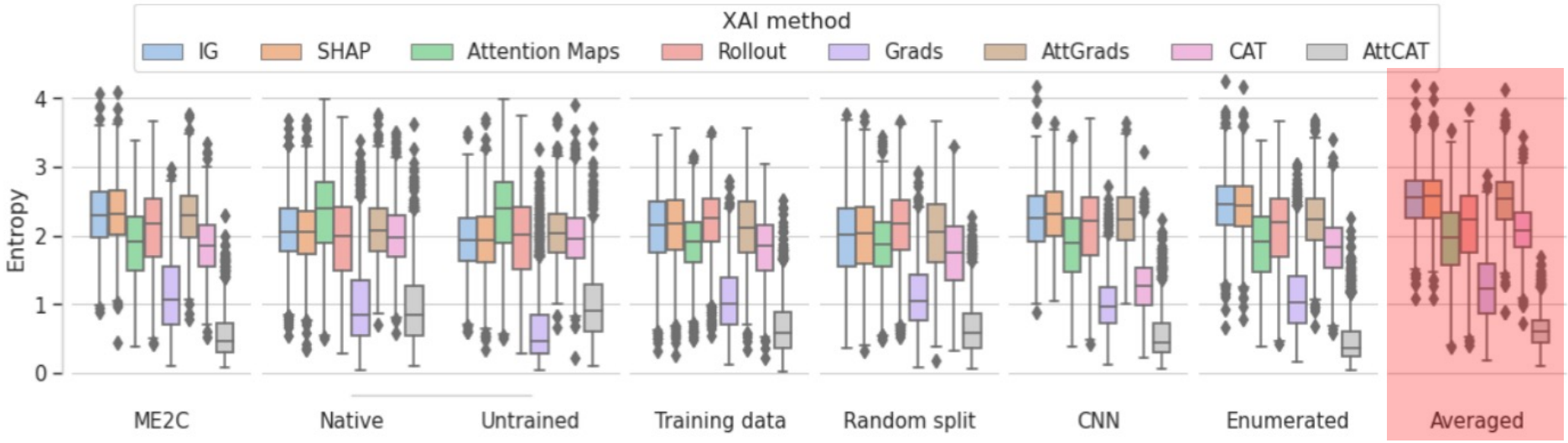


Figure S4: **Entropy values of each XAI method per experiment.** Entropy values of XAI methods of canonical representations or indicating relative information in the attributions. Averaged experiment variation used the average over all enumerations instead of the canonical representation for it's entropy analysis.

Relative Importance Structural Alerts

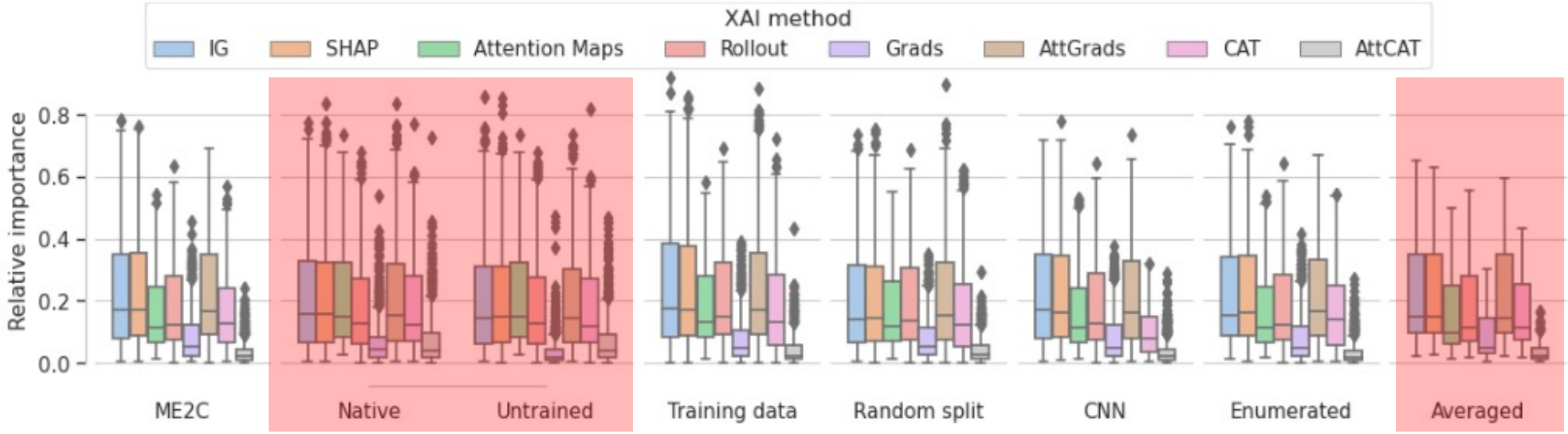


Figure 5: **Relative importance given to expert-derived structural alerts.**

Summary and Conclusions

- **Different representations do not overlap**
  - In-domain has more variance than out-of-domain
  - Randomized models have similar variance as out-of-domain

  ➢*XAI methods for NLP depend mostly on tokenization, not learned parameters*

- **Importance of structural alerts does not indicate importance**
  - Relative importance stays consistent even in completely random models

  ➢*Human intuition should not be used to validate XAI*

- **Test-time augmentation**
  - can be used as a measure of robustness, but needs additional analyses
  - is usefull for increasing agreement between methods and models

Acknowledgments

BSc and MSc students:

Andrea Hunklinger, Fabian Krüger, Isak Palenius, Erik Hansson

Supervisory Committee

Molecular AI Department, Astrazeneca

- *Alessandro Tibo*, Annie Westerlund

The AIDD ESRs

- *Emma Svensson*: Review & Feedback paper
- *Paula Torren Peraire*: Beamsearch Implementation

AstraZeneca

HELMHOLTZ MUNICH

Dr. Igor Tetko

Dr. Samuel Genheden

Prof. Dr. Dr. Fabien Theis

Advanced machine learning for Innovative Drug Discovery (AIDD)
Horizon 2020 Marie Skłodowska-Curie Innovative Training Network - European Industrial Doctorate

All figures in this presentation were created with BioRender.com

The AIDD Team

Early-Stage Researchers (ESRs)

ESR1

Peter Hartog

ESR2

Emma Svensson

ESR3

Paula Torren Peraire

ESR4

Varvara Voinarovska

ESR5

Julian Cremer

ESR6

Son Hà

ESR7

Alan Kai Hassen

ESR8

Ana Sanchez

ESR9

Yasmine Nahal

ESR10

Rosa Friesacher

ESR11

Vincenzo Pamacci

ESR12

Mikhail Andronov

ESR13

Allesio Fallani

ESR14

Muhammad Arslan Masood

ESR15

Mathias Hilfiker

ESR16

Mariia Radaeva

Any Questions?

Thank you for your attention!

- **Different representations do not overlap**
  - ➢ *XAI methods for NLP depend mostly on tokenization, not learned parameters*
- **Importance of structural alerts does not indicate importance**
  - ➢ *Human intuition should not be used to validate XAI*
- **Test-time augmentation can be used as a measure of robustness**
  - Needs additional analyses

References

(1) Jannis Born, Greta Markert, Nikita Janakarajan, Talia B Kimber, Andrea Volkamer, María Rodríguez Martínez, and Matteo Manica. Chemical representation learning for toxicity prediction. Digital Discovery, 2023.

(2) Pavel Karpov, Guillaume Godin, and Igor V Tetko. Transformer-cnn: Swiss knife for qsar modeling and interpretation. Journal of cheminformatics, 12(1):1–12, 2020.

(3) Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. " why should i trust you?" explaining the predictions of any classifier. In Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining, pages 1135–1144, 2016.

(4) Mukund Sundararajan, Ankur Taly, and Qiqi Yan. Axiomatic attribution for deep networks, 2017.

(5) Yao Qiang, Deng Pan, Chengyin Li, Xin Li, Rhongho Jang, and Dongxiao Zhu. Attcat: Explaining trans-formers via attentive class activation tokens. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho,and A. Oh, editors, Advances in Neural Information Processing Systems, volume 35, pages 5052–5064. Cur-ran Associates, Inc., 2022. URL https://proceedings.neurips.cc/paper_files/paper/2022/file/20e45668fefa793bd9f2edf19be12c4b-Paper-Conference.pdf.

(6) Kexin Huang, Tianfan Fu, Wenhao Gao, Yue Zhao, Yusuf Roohani, Jure Leskovec, Connor W Coley, Cao Xiao, Jimeng Sun, and Marinka Zitnik. Therapeutics data commons: Machine learning datasets and tasks for drug discovery and development. arXiv preprint arXiv:2102.09548, 2021.

(7) Jeroen Kazius, Ross McGuire, and Roberta Bursi. Derivation and validation of toxicophores for mutagenicity prediction. Journal of medicinal chemistry, 48(1):312–320, 2005.

(8) Irwin, R., Dimitriadis, S., He, J. & Bjerrum, E. J. Chemformer: a pre-trained transformer for computational chemistry. Mach Learn Sci Technol 3, 015022 (2022).

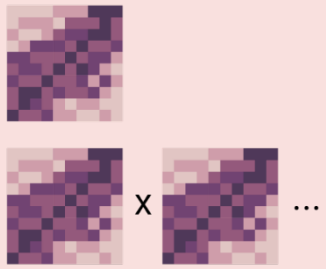
(9) Hinton, G., Vinyals, O. & Dean, J. Distilling the knowledge in a neural network. arXiv preprint arXiv:1503.02531 (2015).

(10) Shazeer, N. et al. Outrageously large neural networks: The sparsely-gated mixture-of-experts layer. arXiv preprint arXiv:1701.06538 (2017).

(11) Hunklinger, Andrea, et al. "The openOCHEM consensus model is the best-performing open-source predictive model in the First EUOS/SLAS Joint Compound Solubility Challenge." SLAS Discovery (2024).

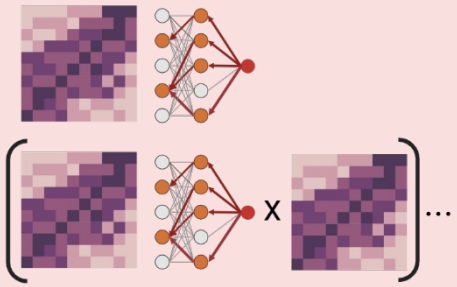
Attention Maps

Rollout



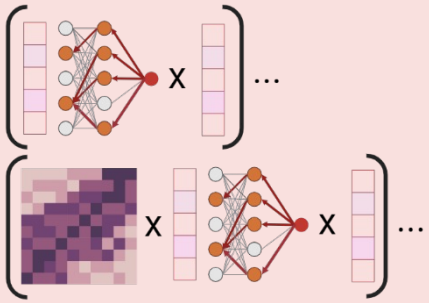
Grads

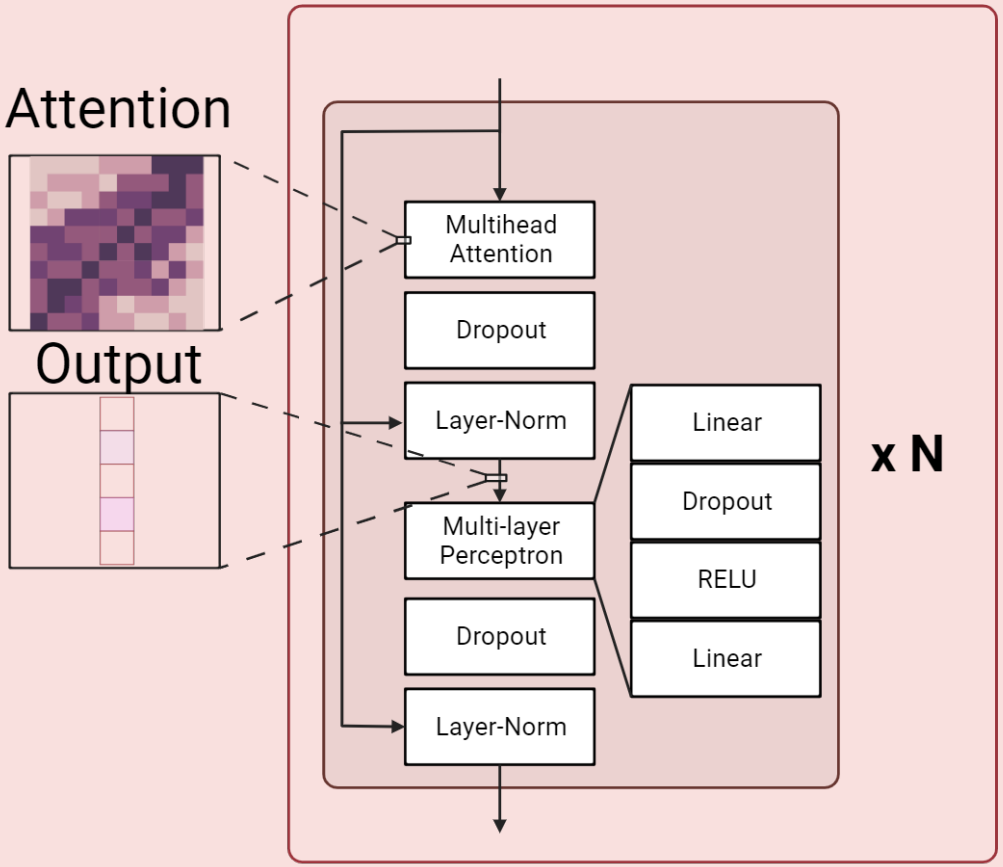
AttGrads



CAT

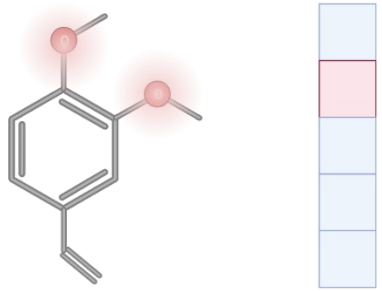
AttCAT





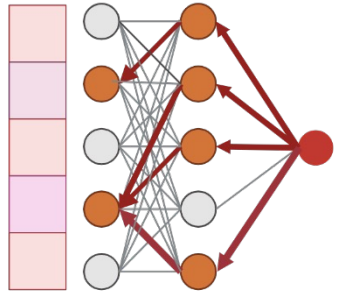
XAI Methods

Pertubation-based XAI



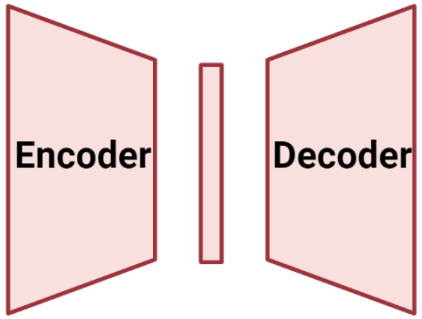
SHAP[3]

Gradient-based XAI



Integrated Gradients (IG)[4]

Transformer-dependent XAI



Attention Maps, Rollout, Grads, AttGrads, CAT and AttCAT[5]

(3) Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. " why should i trust you?" explaining the predictions of any classifier. In Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining, pages 1135–1144, 2016.
(4) Mukund Sundararajan, Ankur Taly, and Qiqi Yan. Axiomatic attribution for deep networks, 2017.
(5) Yao Qiang, Deng Pan, Chengyin Li, Xin Li, Rhongho Jang, and Dongxiao Zhu. Attcat: Explaining trans-formers via attentive class activation tokens. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho,and A. Oh, editors, Advances in Neural Information Processing Systems, volume 35, pages 5052–5064. Cur-ran Associates, Inc., 2022. URL https://proceedings.neurips.cc/paper_files/paper/2022/file/20e45668fefa793bd9f2edf19be12c4b-Paper-Conference.pdf.

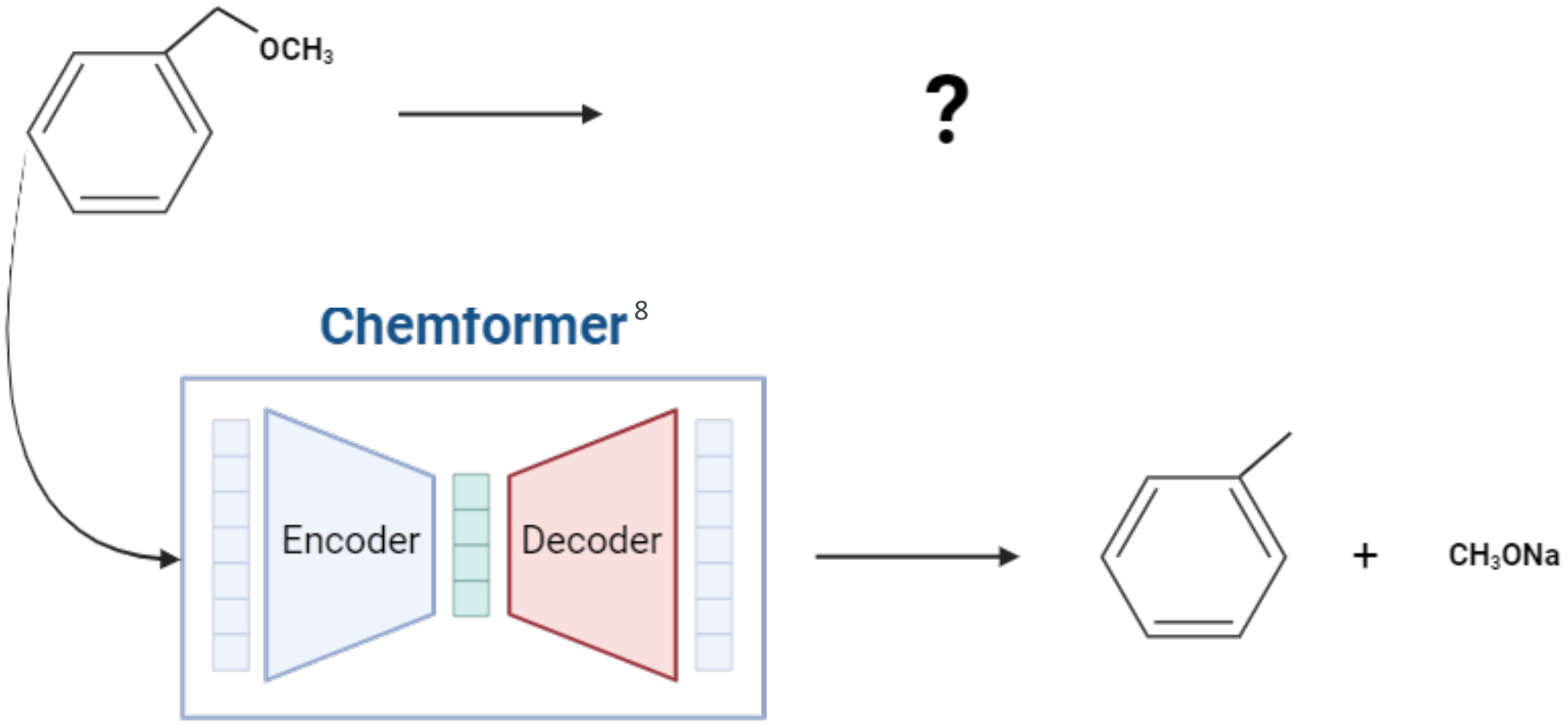
AstraZeneca

Other Projects

P.B.R. Hartog

22 april 2024

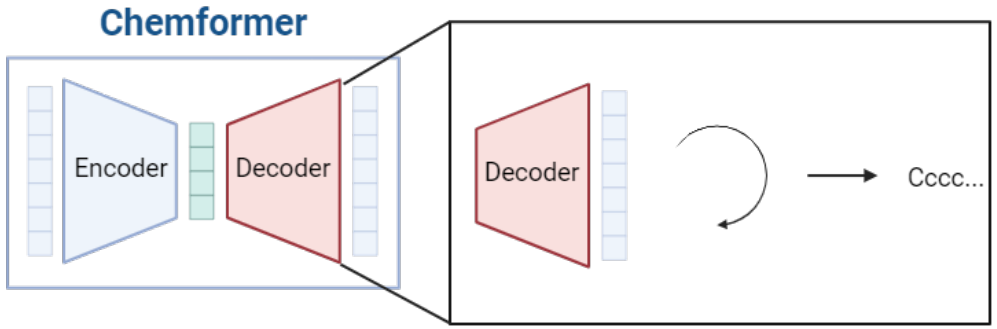
Retrosynthesis

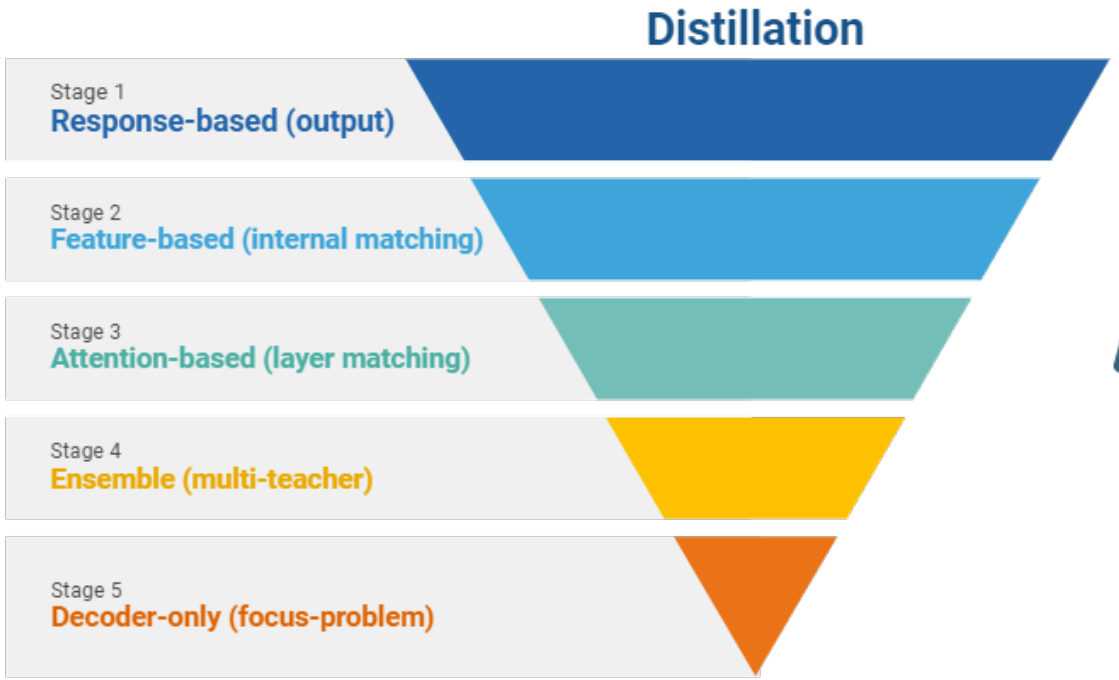


8) Irwin, R., Dimitriadis, S., He, J. & Bjerrum, E. J. Chemformer: a pre-trained transformer for computational chemistry. Mach Learn Sci Technol 3, 015022 (2022).

Chemformer Distillation

- Beam-search
  - Too slow
  - Too energy intensive
- Distillation: bigger ≠ better[9]
  - Use a large model to train a smaller model with richer information





9) Hinton, G., Vinyals, O. & Dean, J. Distilling the knowledge in a neural network. arXiv preprint arXiv:1503.02531 (2015).

BSc and MSc Projects

- ✓Andrea Hunklinger: TUM BSc Student
  - Using OCHEM and ensembling to win the Kaggle SLAS solubility challenge[11]
- ✓Fabian Krüger: TUM MSc Student
  - Uncertainty Quantification for using smiles enumeration and MC dropout
- Isak Palenius & Erik Hansson: Chalmers MSc Student
  - Using influence functions to explain model predictions

11) Hunklinger, Andrea, et al. "The openOCHEM consensus model is the best-performing open-source predictive model in the First EUOS/SLAS Joint Compound Solubility Challenge." SLAS Discovery (2024).

Finishing the PhD

Doctoral Candidacy

- ✓ TUM Kick-Off Seminar
- ✓ HELENA Kick-off Seminar
- ✓ Registered Thesis Committee
- Annual TAC Meetings
  - ✓ 9 Months
  - ✓ 1.5 years
  - 2.5 years
- 270 hours of Scientific Training
- 40 hours in Professional Skills (flexible)

Thesis Requirements

Minimal Dissertation Requirements

- Active conference participation
- Submitted article

Publication-based Thesis

- Two accepted peer-reviewed articles
  - First author or co-authorship

Scientific Development

Scientific Training

Courses:
- ✓ Fundamentals of Project Management
- ✓ Optimizing Writing Strategies
- ✓ First Steps in Your Doctorate
- ✓ Presenting your Research Competently
- ✓ AIDD Seminars

AIDD Schools:
- ✓ Munich
- ✓ Lugano
- ✓ Leuven
- ✓ AALTO
- ✓ AstraZeneca
- ✓ Berlin

Conferences

- ✓ PhysChem Forum
- ✓ NeurIPS 2022
- ✓ <Interact>
- ✓ BayesComp2023
- • ICANN 2024

Collaboration

- ✓ AIDD Schools
- ✓ Collaboration Emma Svensson
- ✓ Secondment JKU
- ✓ Molecular AI Days
- • NextGen DS AstraZeneca

Ames Results

Table 5: **AUROC, accuracy, F1, MCC precision and recall scores of MLP models transfer learned on Ames data.** Values are based on the scaffold split.

	training	AUROC↑	Accuracy↑	F1↑	MCC↑	Precision↑	Recall↑
no training	Untrained	0.652	0.516	0.143	0.063	0.081	0.619
	Native	0.676	0.634	0.624	0.269	0.607	0.642
variations	Random split	0.856	0.788	0.788	0.576	0.789	0.787
	Train set	0.873	0.792	0.792	0.584	0.792	0.792
	Enumerated	0.772	0.696	0.697	0.393	0.699	0.695
models	CNN	0.709	0.650	0.657	0.300	0.671	0.644
	NN	0.810	0.739	0.739	0.478	0.738	0.739

Augmentation improves agreement

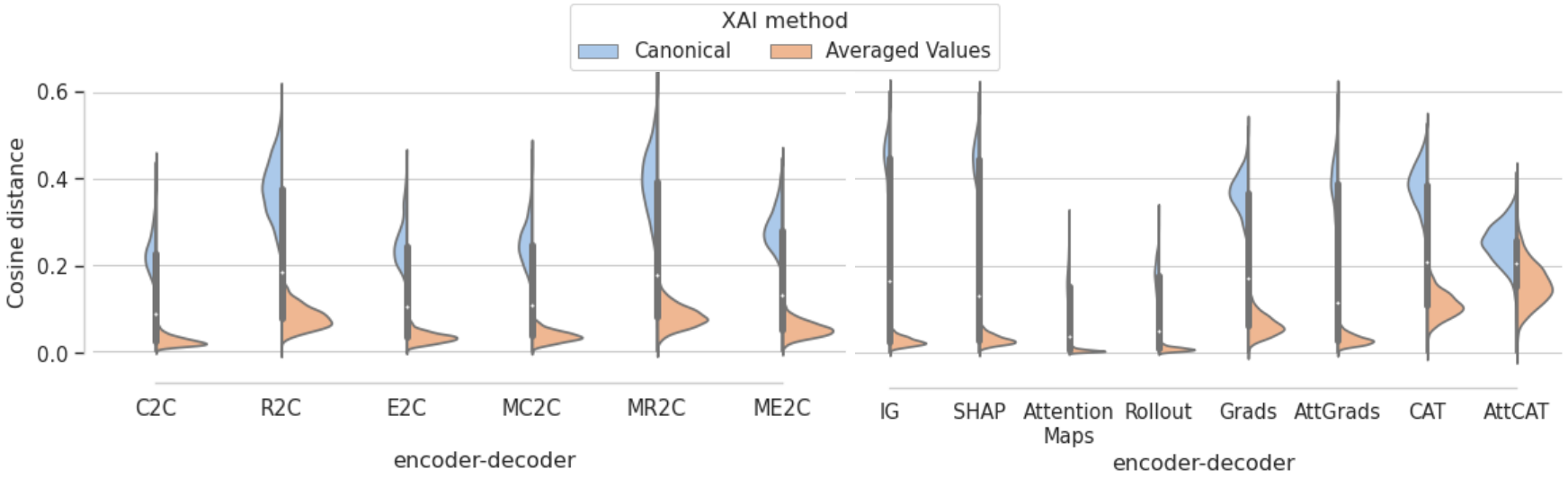


Figure 4: **Comparison of the canonical and averaged values of in-between model and in-between method cosine distances.** Cosine distances between different encoder-decoder models for the same method (in-between model) and different methods for the same encoder-decoder model (in-between method) of canonical and averaged atom attributions.

Dissertation questions

- Theme?
  - Dissecting text-based molecular representation models.
- Publication-based or monograph?
- (Latex) template?

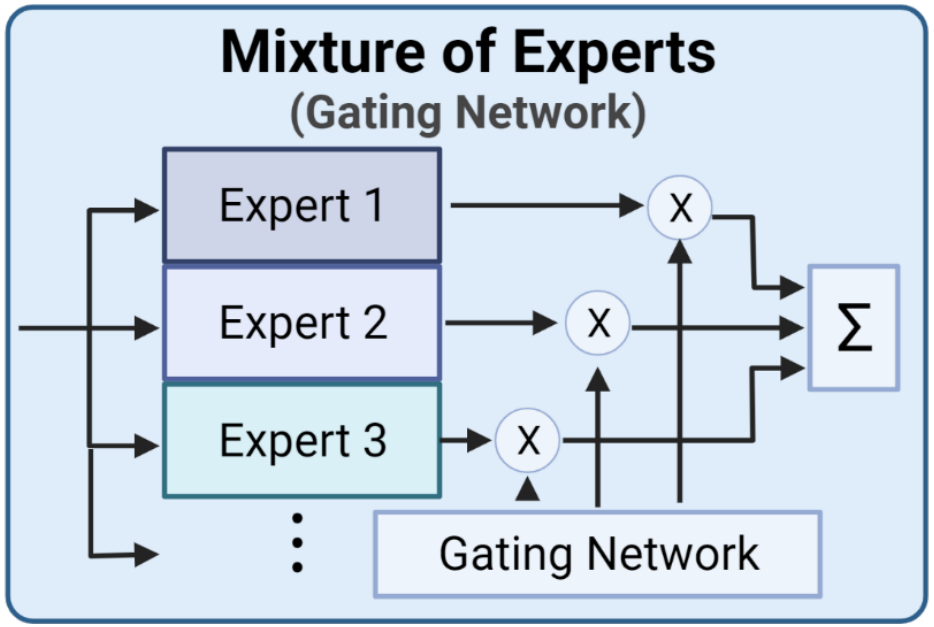
Model Combination using Mixture-of-Experts

Synthesis prediction often fails on ringbreaking reactions

There are two models in AZ

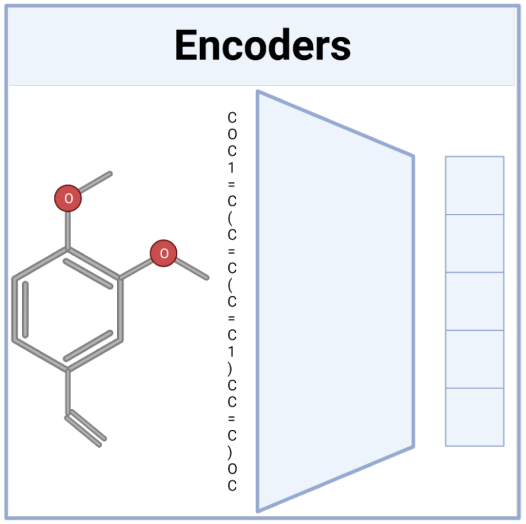
- General model
- Specialized model

➢Combine two synthesis models using model distillation and MoE[10]



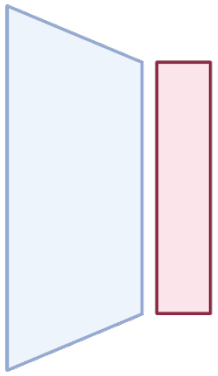
10) Shazeer, N. et al. Outrageously large neural networks: The sparsely-gated mixture-of-experts layer. arXiv preprint arXiv:1701.06538 (2017).

ChEMBL models



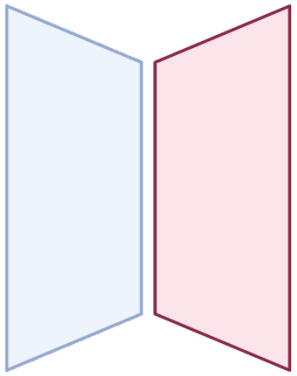
BERT[1]

In: Masked Smiles
Out: Masked Index Values



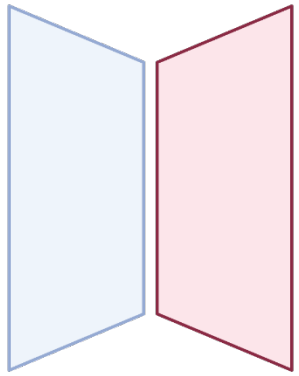
Seq2Seq[2]

In: Smiles
Out: Smiles



BART[3]

In: Masked Smiles
Out: Smiles



1) Devlin, J., Chang, M.-W., Lee, K. & Toutanova, K. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv Prepr. arXiv1810.04805* (2018).
2) Prakash, A. *et al.* Neural paraphrase generation with stacked residual LSTM networks. *arXiv Prepr. arXiv1610.03098* (2016).
3) Lewis, M. *et al.* Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. *arXiv Prepr. arXiv1910.13461* (2019).

ChEMBL Results

Table S1: **Model Statistics of pretrained transformer models.** Bold values are accuracies calculated without previous token information. Models were tested on canonical SMILES to canonical SMILES (canonical) and enumerated SMILES to canonical SMILES (enumerated).

		canonical		**enumerated**					
architecture	training	character accuracy		sequence accuracy		character accuracy		sequence accuracy		time (hh:mm)	size
encoder only	C2C	**1.000**		**1.000**		**1.000**		**1.000**		01:16	6.7M
	R2C	**0.202**		**0.000**		**0.195**		**0.000**		00:42	6.7M
	E2C	**1.000**		**1.000**		**1.000**		**1.000**		03:33	6.7M
	MC2C	**0.999**		**0.993**		**0.994**		**0.819**		01:25	6.7M
	MR2C	**0.202**		**0.000**		**0.196**		**0.000**		00:46	6.7M
	ME2C	**1.000**		**0.991**		**0.999**		**0.962**		10:31	6.7M
encoder-decoder	C2C	0.998	**0.995**	0.994	**0.994**	0.998	**0.996**	0.994	**0.994**	01:45	15.9M
	R2C	0.995	**0.727**	0.923	**0.215**	0.788	**0.205**	0.084	**0.020**	02:44	15.9M
	E2C	0.998	**0.995**	0.995	**0.995**	0.998	**0.995**	0.994	**0.994**	08:43	15.9M
	MC2C	0.998	**0.996**	0.993	**0.987**	0.976	**0.965**	0.463	**0.452**	02:53	15.9M
	MR2C	0.995	**0.525**	0.935	**0.168**	0.794	**0.188**	0.085	**0.015**	05:56	15.9M
	ME2C	0.998	**0.994**	0.983	**0.958**	0.998	**0.994**	0.979	**0.936**	24:04	15.9M

Ames Results

Table 5: AUROC, accuracy, F1, MCC precision and recall scores of MLP models transfer learned on Ames data. Values are based on the scaffold split.

	training	AUROC↑	Accuracy↑	F1↑	MCC↑	Precision↑	Recall↑
no training	Untrained	0.652	0.516	0.143	0.063	0.081	0.619
	Native	0.676	0.634	0.624	0.269	0.607	0.642
variations	Random split	0.856	0.788	0.788	0.576	0.789	0.787
	Train set	0.873	0.792	0.792	0.584	0.792	0.792
	CNN	0.709	0.650	0.657	0.300	0.671	0.644
	Enumerated	0.810	0.739	0.739	0.478	0.738	0.739
encoder only	C2C	0.734	0.666	0.666	0.332	0.665	0.666
	R2C	0.738	0.670	0.670	0.339	0.670	0.670
	E2C	0.731	0.665	0.665	0.331	0.665	0.665
	MC2C	0.754	0.682	0.682	0.364	0.683	0.682
	MR2C	0.694	0.653	0.652	0.305	0.651	0.653
	ME2C	0.804	0.719	0.719	0.438	0.719	0.719
encoder-decoder	C2C	0.716	0.662	0.662	0.324	0.663	0.662
	R2C	0.698	0.634	0.634	0.269	0.634	0.634
	E2C	0.748	0.677	0.678	0.354	0.679	0.676
	MC2C	0.751	0.682	0.682	0.365	0.681	0.683
	MR2C	0.698	0.647	0.647	0.294	0.647	0.647
	ME2C	0.772	0.696	0.697	0.393	0.699	0.695

Random vs Canon

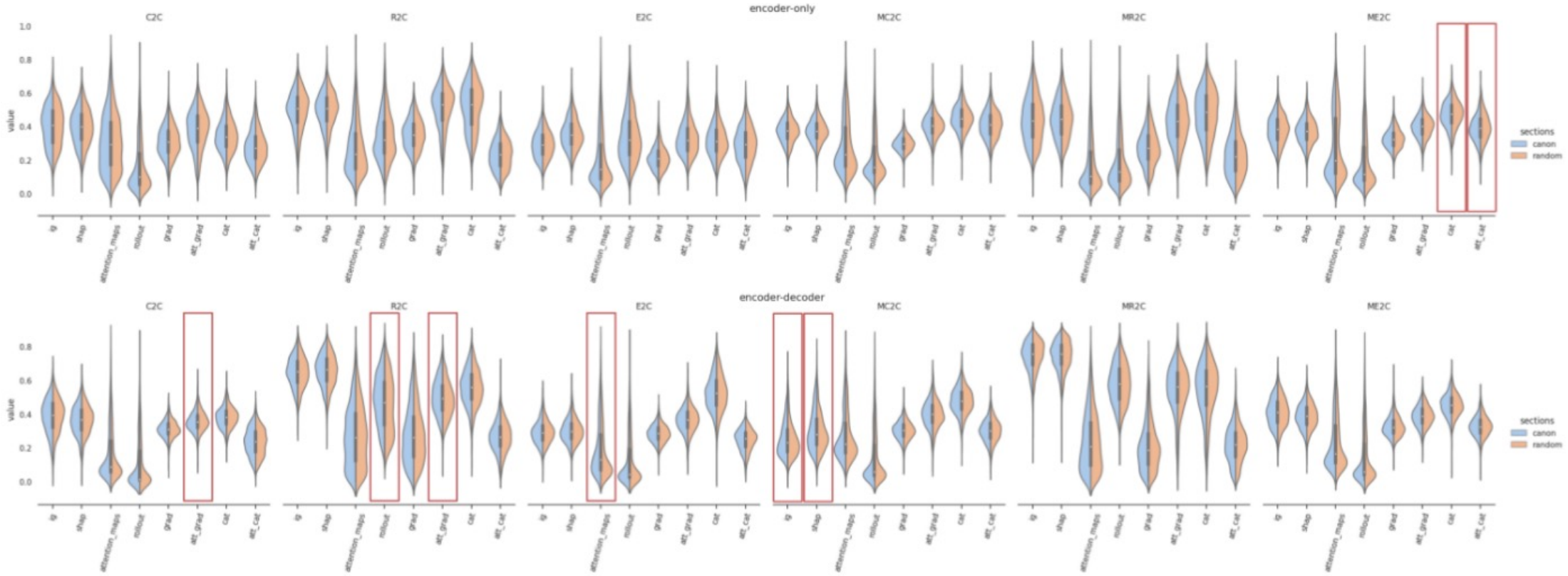


Figure S2: **Violin plot of in-sample cosine distance populations per model and XAI method.** Red boxes indicate p < 0.005 and notes that the populations are not significantly similar.

Random vs Canon (2)

Table S3: **Mann Whitney U test statistics of the difference in in-sample distance populations for each model and interpretation method.** Values are compared between canonical SMILES representation attributions or random SMILES representation as compared to all other enumerated values. Bold values are $p > 0.005$.

architecture	training	IG	SHAP	Att. Maps	Rollout	Grads	AttGrads	CAT	AttCAT
encoder only	C2C	0.761	0.867	0.767	0.722	0.831	0.032	0.930	0.673
	R2C	0.121	0.114	0.507	0.339	0.316	0.036	0.376	0.469
	E2C	0.375	0.796	0.871	0.299	0.667	0.908	0.900	0.081
	MC2C	0.848	0.978	0.370	0.548	0.281	0.342	0.921	0.409
	MR2C	0.711	0.760	0.925	0.793	0.903	0.278	0.457	0.855
	ME2C	0.216	0.373	0.626	0.955	0.326	0.060	**0.004**	**0.000**
encoder-decoder	C2C	0.414	0.202	0.794	0.794	0.537	**0.001**	0.060	0.261
	R2C	0.563	0.324	0.765	**0.000**	0.657	**0.000**	0.237	0.376
	E2C	0.066	0.348	**0.000**	0.046	0.422	0.167	0.814	0.536
	MC2C	**0.004**	**0.002**	0.651	0.932	0.447	0.521	0.625	0.362
	MR2C	0.055	0.398	0.960	0.861	0.668	0.798	0.154	0.356
	ME2C	0.590	0.381	0.955	0.971	0.498	0.186	0.377	0.376

Expert-derived Structural Alerts

toxicophore name	smarts
specific arom nitro	O=N(~O)a
specific arom amine	a[NH2]
aromatic nitroso	a[N;X2]=O
alkyl nitrite	CO[N;X2]=O
nitrosamine	N[N;X2]=O
epoxide	O1[c,C]-[c,C]1
aziridine	C1NC1
azide	N=[N+]=[N-]
diazo	C=[N+]=[N-]
triazene	N=N-N
aromatic azo	c[N;X2]!@;=[N;X2]c
aromatic azoxy	cN!@;=[N;X3](O)c
unsubstituted heteroatom-bonded heteroatom	[OH,NH2][N,O]
hydroperoxide	[OH]O
oxime	[OH][N;X2]=C
1,2-disubstituted peroxide	[c,C]OO[c,C]
1,2-disubstituted aliphatic hydrazine	C[NH][NH]C
aromatic hydroxylamine	[OH]Na
aliphatic hydroxylamine	[OH]N
aromatic hydrazine	[NH2]Na
aliphatic hydrazine	[NH2]N
diazohydroxyl	[OH][N;X2]=[N;X2]
aliphatic halide	[Cl,Br,I]C
carboxylic acid halide	[Cl,Br,I]C=O
nitrogen or sulphur mustard	[N,S]!@[C;X4]!@[CH2][Cl,Br,I]
aliphatic monohalide	[Cl,Br,I][C;X4]

Expert-derived Structural Alerts (2)

alpha-chlorothioalkane	SC[Cl]
beta-halo ethoxy group	[Cl,Br,I]!@[C;X4]!@[C;X4]O
chloroalkene	[Cl]C([X1])=C[X1]
1-chloroethyl	[Cl,Br,I][CH][CH3]
polyhaloalkene	[Cl,Br,I]C(([F,Cl,Br,I])[X1])C=C
polyhalocarbonyl	[Cl,Br,I]C(([F,Cl,Br,I])[X1])C(=O)[c,C]
bay-region in polycyclic aromatic hydrocarbons	[cH]1[cH]ccc2c1c3c(cc2)cc[cH][cH]3
k-region in polycyclic aromatic hydrocarbons	[cH]1cccc2c1[cH][cH]c3c2ccc[cH]3
polycyclic aromatic system	a13~a~a~a~a2~a1~a(~a~a~a~3)~a~a~a~2
polycyclic aromatic system	a1~a~a~a2~a~1~a~a3~a(~a~2)~a~a~a~3
polycyclic aromatic system	a1~a~a~a2~a~1~a~a~a3~a~2~a~a~a~3
polycyclic aromatic system	a1~a~a~a~a2~a~1~a3~a(~a~2)~a~a~a~a~3
polycyclic aromatic system	a1~a~a~a~a2~a~1~a~a3~a(~a~2)~a~a~a~3
polycyclic aromatic system	a1~a~a~a~a2~a~1~a~a3~a(~a~2)~a~a~a~a~3
polycyclic aromatic system	a1~a~a~a~a2~a~1~a~a~a3~a~2~a~a~a~3
polycyclic aromatic system	a1~a~a~a~a2~a~1~a~a~a3~a~2~a~a~a~a~3
polycyclic aromatic system	a13~a~a~a~a2~a1~a(~a~a~a~3)~a~a~2
sulphonate-bonded carbon (alkyl alkane sulphonate or dialkyl sulphate)	[$([C,c]OS((=O)=O)O!@[c,C]),$([c,C]S((=O)=O)O!@[c,C])]
aliphatic N-nitro	O=N(~O)N
alpha,beta unsaturated aldehyde (including alpha-carbonyl aldehyde)	[$(O=[CH]C=C),$(O=[CH]C=O)]
diazonium	[N;v4]#N
beta-propiolactone	O=C1CCO1
alpha,beta unsaturated alkoxy group	[CH]=[CH]O
1-aryl-2-monoalkylhydrazine	[NH;!R][NH;!R]a
aromatic methylamine	[CH3][NH]a
ester derivative of aromatic hydroxylamine (including original specific toxicophore)	aN([$([OH]),$(O*=O)])[$([#1]),$(C(=O)[CH3]),$([CH3]),$([OH]),$(O*=O)]

Component Breakdown

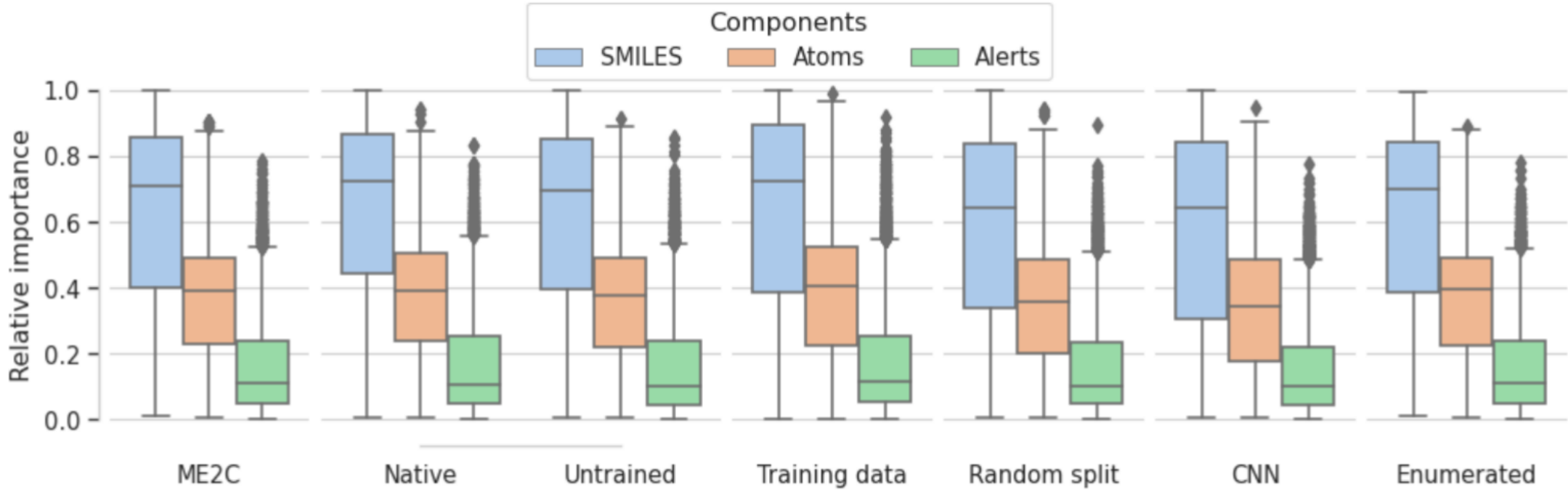
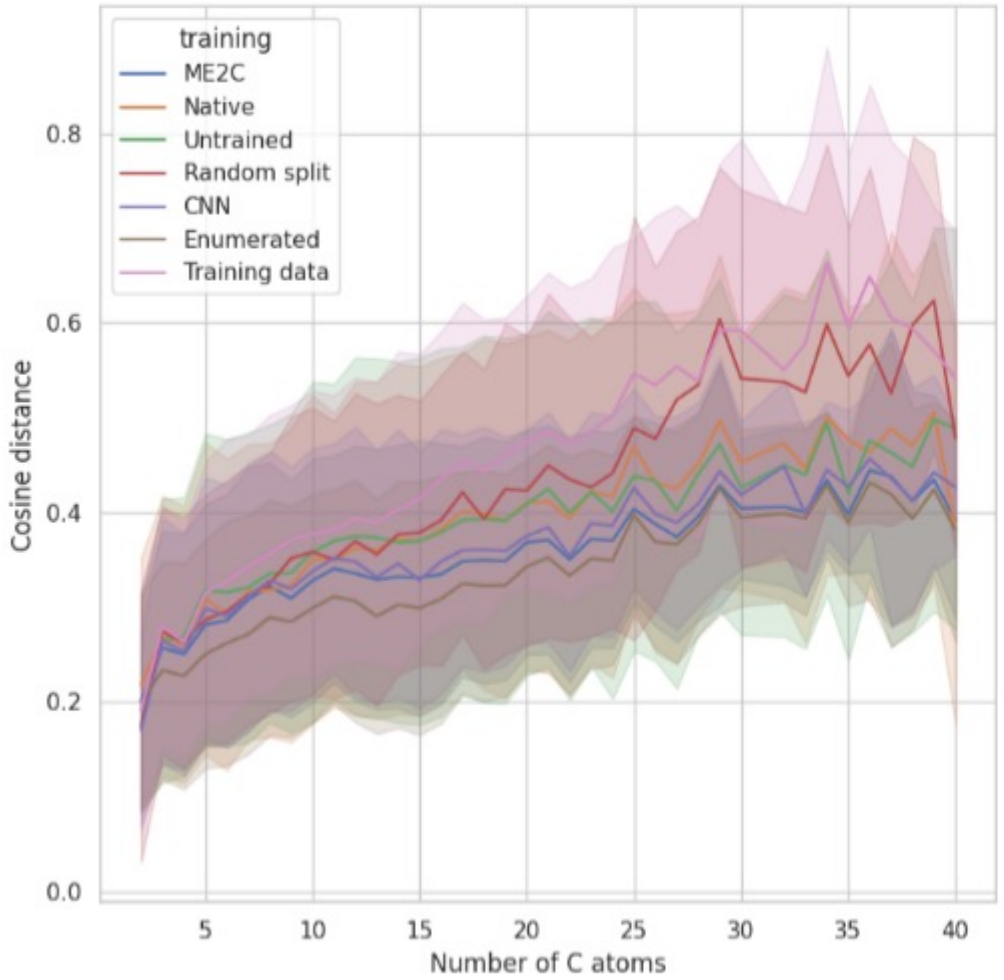


Figure S8: **Relative importance given to different components for experiments.** Relative token importance given to smiles, atoms and atoms corresponding to expert-derived structural alerts. Relative importance is given for canonical representations and aggregated over all XAI methods of different pre-trained representation models.

Carbon breakdown



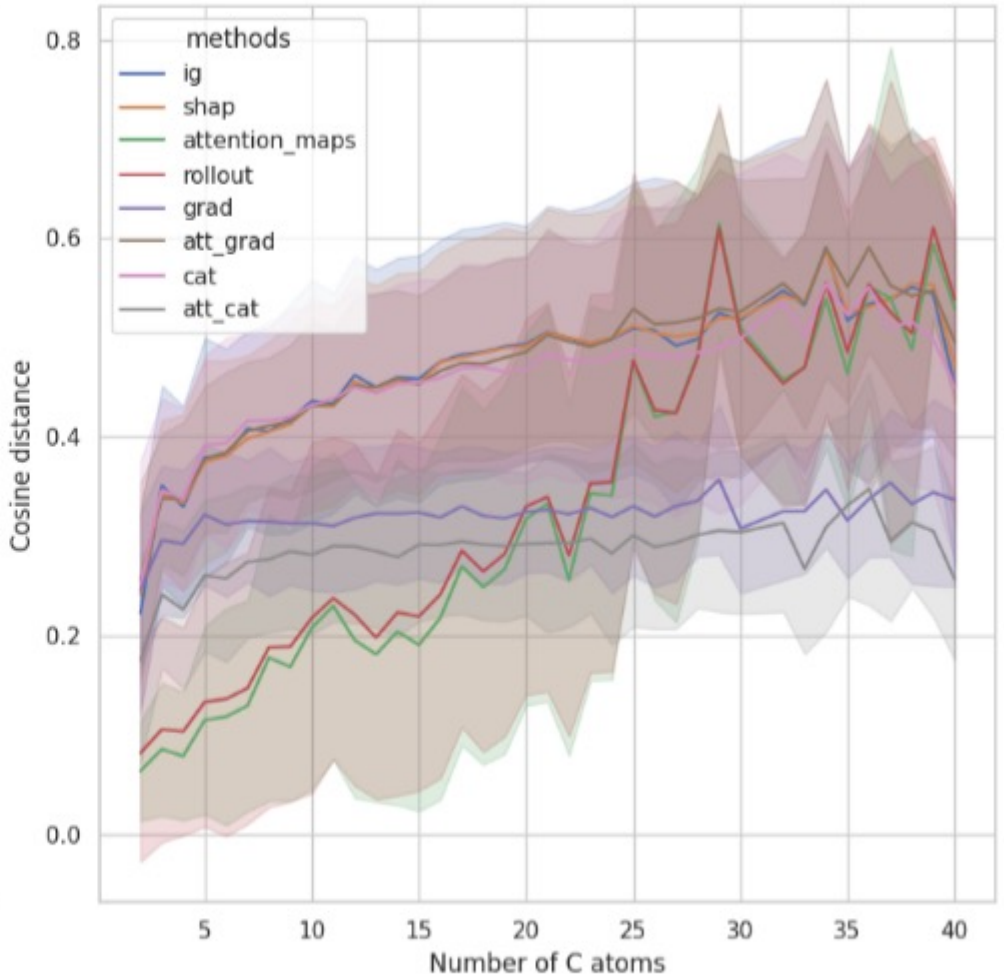


Figure S5: **Cosine distance as a measure of number of carbon atoms.** Variations of in-between sample cosine distances and carbon atoms, broken down to show training variations and XAI methods.

Supporting information