Known Unknowns: Out-of-Distribution Property Prediction in Materials and Molecules

Nofit Segal[1*], **Aviv Netanyahu**[2*], **Kevin P. Greenman**[3,4], **Pulkit Agrawal**[2†], and **Rafael Gómez-Bombarelli**[1†]

[1]Materials Science and Engineering, MIT
[2]Electrical Engineering and Computer Science, MIT
[3]Chemical Engineering, Catholic Institute of Technology
[4]Chemistry, Catholic Institute of Technology
{nofit, avivn, pulkitag, rafagb}@mit.edu, kgreenman@catholic.tech

Abstract

Developing high-performance materials and molecules often requires identifying those with property values that fall outside the known distribution. Therefore, the ability to extrapolate to out-of-support (OOS) property values is critical for both solid-state materials and molecular design. Given the chemical compositions of solids or SMILES of molecules and their property values, our objective is to learn a predictor that extrapolates zero-shot to higher ranges. In this work, we employ a transductive approach to property prediction and achieve more accurate predictions, as well as a 3x and 2.5x improvement in True Positive Rate (TPR) of OOS materials and molecules identification, respectively. We leverage analogical input-target relations in the training and test sets, enabling generalization beyond the training target support.

1 Introduction

Designing new materials and molecules is essential for the development of new technologies. Traditionally, this design process involves extensive experimental trial and error or high-throughput simulations to screen databases, both of which are time-consuming and resource-intensive [1, 2]. As a result, there is increasing interest in applying machine learning (ML) techniques to accelerate the discovery of materials and molecules with desired properties [1–5].

One strategy for finding materials with desired properties is inverse design through conditional generation [4–8, 2]. A complementary approach is to screen candidate materials and molecules through property prediction [9–14]. However, both approaches typically struggle when property values fall outside the training distribution [4, 11, 13, 15, 16]. Enhancing extrapolative capabilities in property prediction would improve large dataset screening by identifying promising compounds and molecules with exceptional properties. This approach could help guide further synthesis and computational efforts, ultimately advancing materials and molecular design.

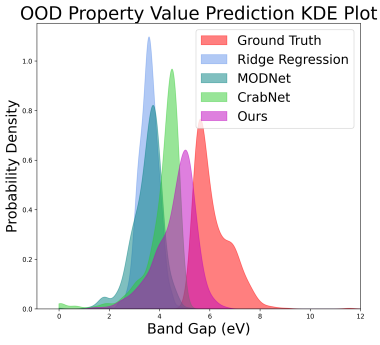


Figure 1: ML methods often fail in out-of-support prediction. Our transductive approach predicts values closer to the desired distribution.

*Denotes equal contribution. † Denotes equal advising.

AI4Mat Workshop at the 38th Conference on Neural Information Processing Systems (NeurIPS 2024).

Bilinear Transduction [17] has been successful in extrapolation in robotics domains w.r.t. the input space by making predictions based on analogies. In this work, we empirically investigate whether Bilinear Transduction improves extrapolation in materials science (Figure 1). Specifically, extrapolation in the output space, i.e. the property value, for finding high-performance materials.

Our main contributions are (1) adapting a method for extrapolation to a new application in materials science and (2) an extensive evaluation of our approach on common high-throughput calculated and experimental solid-state materials and molecules benchmarks, showcasing improved extrapolation to out-of-support (OOS).

2 Method

2.1 Bilinear Transduction Preliminaries

Bilinear Transduction [17] is a regression method for OOS generalization, a type of distribution shift where data becomes OOS at test time. By reparameterizing the problem, OOS generalization is possible via a bilinear predictor. Let $\mathcal{X} \subseteq \mathbb{R}^n$ denote a representation space of data points and let $\mathcal{Y} \subseteq \mathbb{R}$ denote the space of labels. Given a training set $\mathcal{D}^{\text{tr}} = \{(x, y)\}_n \subseteq \mathcal{X} \times \mathcal{Y}$, Bilinear Transduction learns a transductive predictor that extrapolates in a zero-shot manner to OOS test points. Naively, one can learn predictor $h_\theta : \mathcal{X} \to \mathcal{Y}$. However, machine learning methods often fail under covariate shift [18, 19]. Instead, Bilinear Transduction offers the following formulation.

Let $\Delta\mathcal{X} = \{x_i - x_j | x_i, x_j \in \mathcal{X}\}$ denote the differences distribution. During training, predictor $h_\theta : \Delta\mathcal{X} \times \mathcal{X} \to \mathcal{Y}$ predicts value $y_2 \in \mathcal{D}^{\text{tr}}_{\mathcal{Y}}$ for data point $x_2 \in \mathcal{D}^{\text{tr}}_{\mathcal{X}}$ given anchor point $x_1 \in \mathcal{D}^{\text{tr}}_{\mathcal{X}}$ and the difference between them $\Delta x_{21} = x_2 - x_1 \in \mathcal{D}^{\text{tr}}_{\Delta\mathcal{X}}$. The predictor $h_\theta(\Delta x, x) = f_\theta(\Delta x) g_\theta(x)$ is implemented as a bilinear function in non-linear embeddings of Δx and x.

The test set $\mathcal{D}^{\text{te}} \subseteq \mathcal{X}$ includes OOS data points. Given test point $x_{\text{te}} \in \mathcal{D}^{\text{te}}$, its predicted value is $h_\theta(\Delta x_{\text{te,an}}, x_{\text{an}})$. Anchor $x_{\text{an}} \in \mathcal{D}^{\text{tr}}_{\mathcal{X}}$ is the training point that minimizes the distance between its difference with the test point $\Delta x_{\text{te,an}} = x_{\text{te}} - x_{\text{an}}$, and differences within the training distribution $\mathcal{D}^{\text{tr}}_{\Delta X}$. Formally, $x_{\text{an}} = \text{argmin}_{x_i \in \mathcal{D}^{\text{tr}}} \{||\Delta x_{\text{te},i} - \Delta x||_2 \mid \Delta x \in \mathcal{D}^{\text{tr}}_{\Delta X}\}$. This reparameterization converts the problem to within support, as $\Delta x_{\text{te,an}}$ and x_{an} are within the training distribution.

2.2 Bilinear Transduction for Materials

We use descriptor-based representations: composition-based descriptors derived from elemental properties for solids [20] and RDKit [21] descriptors derived from SMILES [22] for molecules. In both cases, using fixed descriptor-based representations offers interpretable features, which are readily found in the periodic table and existing databases [14]. Additionally, for solids, a composition-based approach enables more robust predictions, as it implies weaker assumptions about the material [23]. In our evaluation of materials datasets, we test OOS extrapolation in $\mathcal{Y}$. We split our datasets such that OOS test materials are 5% of data with the highest $\mathcal{Y}$ values. In addition, we randomly sample 5% of the training data for in-distribution evaluation. During training (Algorithm 1), we predict the property value y_i of material x_i from material x_j and their difference $x_i - x_j$, where material x_j has a lower property value.

Algorithm 1 Bilinear Transduction for Materials

1: **Input:** Training set $(x_1, y_1), \ldots, (x_n, y_n)$
2: **Train:** Train θ on loss $\mathcal{L}(\theta) = \sum_{i=1}^{n} \sum_{j:y_j<y_i} \ell(h_\theta(x_i - x_j, x_j), y_i)$
3: **Test:** For each new x_{te}, let $x_{\text{an}} = \text{argmin}_{x_{\text{an}} \in \mathcal{D}^{\text{tr}}_{\mathcal{X}}} \{\|x_{\text{te}} - x_{\text{an}} - \Delta x_{\text{tr}}\|_2 \mid \Delta x_{\text{tr}} \in \mathcal{D}^{\text{tr}}_{\Delta X}\}$, and predict

$$y = h_\theta(x_{\text{te}} - x_{\text{an}}, x_{\text{an}})$$

Bilinear Transduction has theoretical guarantees for extrapolation in OOS input space $\mathcal{X}$ [17]. In our setting, while the model is tested on OOS target values, it does not necessarily operate outside the training input support, and the conditions for the theoretical guarantees are not necessarily met. Descriptor-based features encapsulate fundamental chemical and physical information that directly influences materials and molecular characteristics. Therefore, the difference between feature vectors is related, possibly intricately, to the change in property value. Bilinear Transduction has the potential to extrapolate by learning *how* property values change as a function of compositional differences

instead of predicting these values directly. In this work, we empirically investigate to what degree Bilinear Transduction extends predictions beyond the training target support.

3 Related Work

In the context of materials science, extrapolation can be done with respect to the materials space or the properties space. The first includes generalization to out-of-distribution (OOD) materials structures or chemical spaces. The latter includes extrapolation to OOS property values, as done in this work.

Extrapolation in property space. There has been work on extrapolation to OOS property prediction by encoding nonlinearities of input-to-target relationships into the input material representation [24], engineering interpretable descriptors [25, 26] and applying data imputers [27]. The limitations of classical ML approaches in extrapolating property predictions via regression have also been studied, leading to a shift towards classifying OOS materials instead [15, 16].

Extrapolation in materials space. Recent studies have been focused on developing deep generative models which are suggested to achieve unprecedented levels of OOD generalization towards unseen materials that are dissimilar to the training data, e.g. generalizing to structures with a larger number of atoms or different elemental combinations [28–30]. However, Li et al. [31] argue that in many tasks where OOD is defined with respect to the input materials space, the test sets often fall within the training representation space, making these tasks effectively interpolation rather than extrapolation.

4 Results

We demonstrate Bilinear Transduction's extrapolation capabilities on three common solid materials benchmarks (Section 4.1.1) and one common molecules benchmark (Section 4.2.1), compared against three strong solids baselines (Section 4.1.2) and three strong molecules baselines (Section 4.2.2). Table 1 compares the mean average error (MAE) for OOS predictions on solids and molecules. We include additional results in Appendix A.1. Bilinear Transduction performs consistently better or is comparable to the baselines across differently curated datasets and properties. See Appendix A.3.1 for data representation and processing details for each method.

4.1 Solids

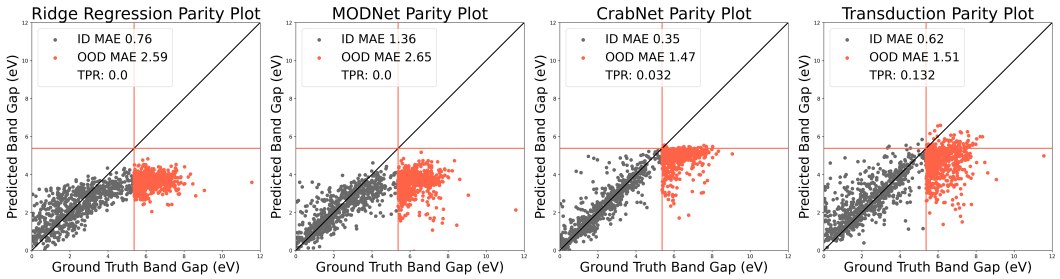
Figure 1 demonstrates that Bilinear Transduction produces a prediction for band gap distribution that is closer to the OOS ground truth distribution. Figure 2 shows that Bilinear Transduction extrapolates to some extent, whereas the other baselines do not exceed the training support threshold.

4.1.1 Datasets

Solids datasets include material compositions and their property values. **AFLOW** contains material property values obtained from high-throughput calculations [32]. Following Kauwe et al. [15], who evaluate classical ML algorithms on AFLOW, we curate a subset of six properties: band gap, bulk modulus, debye temperature, shear modulus, thermal conductivity, and thermal expansion, out of which the last four are scaled by applying a base 10 logarithm. **Matbench** is an automated leaderboard for benchmarking ML algorithms predicting solid material properties [10]. Matbench contains three composition-based regression tasks: experimentally measured band gap [12], experimentally measured yield strength of steels [33], and calculated refractive index [34]. **Materials Project (MP)** provides materials and their property values derived from high-throughput calculations [35]. Following Wang et al. [11], we focus on bulk modulus, shear modulus, and ratio of elastic anisotropy. In cases of duplicate compositions, we retain the entry with the target value corresponding to the lowest formation enthalpy.

4.1.2 Baselines

We compare with **Ridge Regression**, the strongest method in Kauwe et al. [15], who evaluate classical ML algorithms on OOS property values. We further compare with **MODNet** [13] and **CrabNet** [11], leading models in composition-based property prediction.



(a) (b) (c) (d)

Figure 2: **In-distribution, and out-of-support band gap predictions vs. ground truth values**. While (a) Ridge Regression [15], (b) MODNet [13], (c) CrabNet [11] and (d) Bilinear Transduction (ours), perform well within the training distribution (gray dots bounded by the red horizontal line), only Bilinear Transduction extends predictions beyond this range on OOS data (red dots).

4.2 Molecules

Figure 4a demonstrates that Bilinear Transduction produces a prediction for the Freesolv data distribution that is closer to the OOS ground truth distribution. Figures 4b, 4c, 4d and 4e show that Bilinear Transduction extrapolates, whereas the other baselines rarely surpass the boundary marking the beginning of the test support.

4.2.1 Datasets

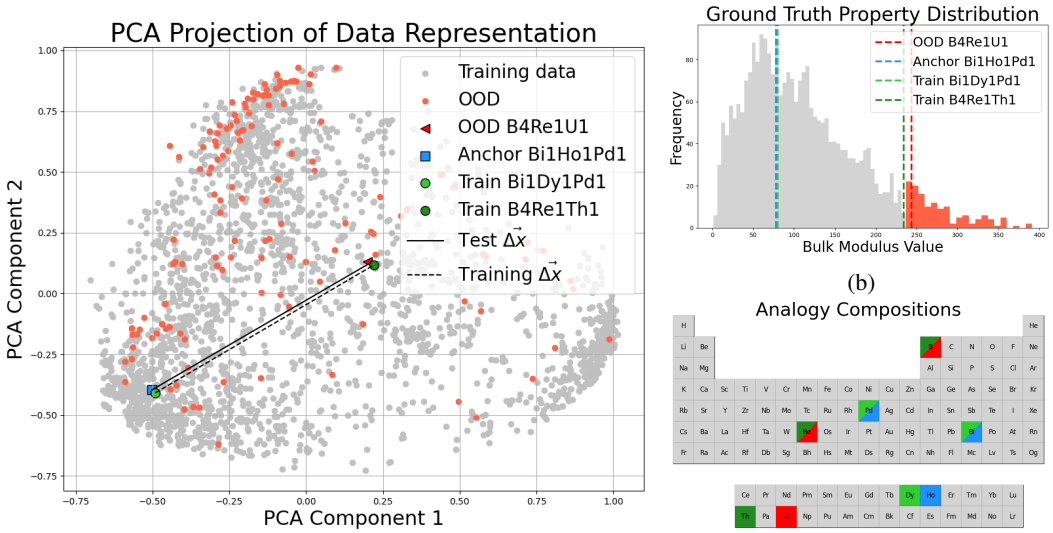
MoleculeNet includes SMILES representations [22] and their property values derived from high-throughput calculations and experimental trials [36]. We focus on physical chemistry and biophysics properties suitable for regression – ESOL, freesolv, lipophilicity and BACE binding.

4.2.2 Baselines

we compare with **Chemprop** [37], a leading method for property prediction from molecular graphs via message-passing. Chemprop has an advantage as its representations include structural information that is not explicitly available in the representation we use – RDKit descriptors [21]. In addition, we compare with **Random Forest (RF)** [38], a classical ML tree-based method, and **Multi Layer Perceptron (MLP)**. These serve as ablations of our method, using the same representation with partial structural information as we do.

Table 1: **Solids (top) and molecules (bottom)** OOS mean average prediction error and standard error of the mean.

Dataset	Property	#Samples	Ridge Reg. [15]	MODNet [13]	CrabNet [11]	Ours
AFLOW [32]	Band Gap [eV]	14123	2.59 ± 0.03	2.65 ± 0.04	$\mathbf{1.47 \pm 0.03}$	1.51 ± 0.04
	Bulk Modulus [GPa]	2740	74.0 ± 3.8	93.06 ± 3.7	59.25 ± 3.2	$\mathbf{47.4 \pm 3.4}$
	Debye Temperature [K]	2740	0.45 ± 0.03	0.62 ± 0.03	0.38 ± 0.02	$\mathbf{0.31 \pm 0.02}$
	Shear Modulus [GPa]	2740	0.69 ± 0.03	0.78 ± 0.04	0.55 ± 0.02	$\mathbf{0.42 \pm 0.02}$
	Thermal Conductivity [$\frac{W}{mK}$]	2734	1.07 ± 0.05	1.5 ± 0.05	0.97 ± 0.03	$\mathbf{0.83 \pm 0.04}$
	Thermal Expansion [K^{-1}]	2733	0.44 ± 0.02	0.47 ± 0.02	$\mathbf{0.37 \pm 0.02}$	0.39 ± 0.02
Matbench [10]	Band Gap [eV]	2154	6.37 ± 0.28	3.26 ± 0.13	2.70 ± 0.13	$\mathbf{2.54 \pm 0.16}$
	Refractive Index	4764	14.4 ± 2.0	4.24 ± 0.48	3.92 ± 0.5	$\mathbf{3.81 \pm 0.49}$
	Yield Strength [MPa]	312	972 ± 34	731 ± 82	740 ± 49	$\mathbf{591 \pm 62}$
MP [35]	Bulk Modulus [GPa]	6307	151 ± 14	60.1 ± 3.9	57.8 ± 4.2	$\mathbf{45.8 \pm 3.9}$
	Elastic Anisotropy	6331	165 ± 17	60.0 ± 4.5	61.4 ± 4.6	$\mathbf{59.8 \pm 4.5}$
	Shear Modulus [GPa]	6184	134.5 ± 7.2	65.6 ± 2.5	65.3 ± 2.8	$\mathbf{63.2 \pm 2.6}$
			Chemprop [37]	**Random Forests** [38]	**MLP** [39]	
MoleculeNet [36]	ESOL [$\frac{mol}{L}$]	1128	0.47 ± 0.04	0.67 ± 0.04	0.5 ± 0.04	$\mathbf{0.42 \pm 0.04}$
	Freesolv [$\frac{kJ}{mol}$]	643	0.44 ± 0.03	0.42 ± 0.02	0.5 ± 0.02	$\mathbf{0.08 \pm 0.01}$
	Lipophilicity [$\log D$]	4200	0.75 ± 0.02	1.02 ± 0.02	0.9 ± 0.02	$\mathbf{0.7 \pm 0.02}$
	BACE binding [IC50]	1513	1.03 ± 0.06	0.93 ± 0.05	0.95 ± 0.07	$\mathbf{0.73 \pm 0.05}$



(a) (b) (c)

Figure 3: **Bilinear Transduction Analogies Visualization**. AFLOW bulk modulus OOS predictions are based on differences between in-distribution anchors and OOS targets, that form analogies to training pairs. (a) OOS-anchor and training pair differences. (b) Ground truth training (gray) and test (red) distributions and OOS, anchor, and analogous training pair values. (c) Analogous compounds. OOS and training target differ by one neighboring f-block element. So do anchor and training anchor.

4.3 Learning Using Analogies

The success of transduction in these tasks is related to one of the tenets of chemistry: *similar materials have similar properties*. The transductive approach effectively builds on this idea, by proposing that *similar changes in chemical compounds or molecular structure* imply *similar changes in properties*. At inference, our approach selects the anchor by minimizing the difference between $\Delta x_{\text{te,an}}$ and Δx_{tr}. We demonstrate how these algebraic operations in the embedding space relate to chemical changes as measured in the domain.

For solids, these operations are expressed as elemental changes. Figure 3 demonstrates this in bulk modulus inference of an OOS sample with stoichiometry B_4ReU. The model selects in-distribution BiHoPd as the anchor, analogous to training anchor BiDyPd and training target B_4ReTh. Specifically, the compositions of these anchors (BiDyPd and BiHoPd), and the targets (B_4ReTh and B_4ReU), differ by only one neighboring f-block element: Dy (Z=66) to Ho (Z=67), and Th (Z=90) to U (Z=92). See Appendix A.2.1 for more examples.

For molecules, these operations manifest as structural similarity. This is notable given that the RDKit descriptor vector used as input lacks detailed structural and connectivity information from the SMILES representation. Figure 7 illustrates this in ESOL inference, with the maximum common structure (MCS) highlighted (Figure 7c) between the anchor-OOS pair, and the training anchor-target pair. Each pair's structures show high similarity, differing by the addition of a conjugated double bond that extends the molecular backbone. See Appendix A.2.2 for additional examples.

5 Discussion

In this work, we demonstrate that Bilinear Transduction improves OOS material and molecular property value prediction both in support coverage and in mean average error. While this is a promising avenue to continue exploring, there are several limitations to this approach. Under our current problem formulation where $\mathcal{Y}$ is OOS, the theoretical guarantees for Bilinear Transduction may not hold. In the future, we plan to (1) investigate why and when the current framework works well on OOS $\mathcal{Y}$ and (2) experiment with learning data representations to fit the Bilinear Transduction assumptions on OOS $\mathcal{X}$. We believe that advancements in OOS property prediction will accelerate the screening of promising compositions and molecules for novel functional materials design.

References

[1] Simon Axelrod, Daniel Schwalbe-Koda, Somesh Mohapatra, James Damewood, Kevin P Greenman, and Rafael Gómez-Bombarelli. Learning matter: Materials design with machine learning and atomistic simulations. *Accounts of Materials Research*, 3(3):343–357, 2022.

[2] Benjamin Sanchez-Lengeling and Alán Aspuru-Guzik. Inverse molecular design using machine learning: Generative models for matter engineering. *Science*, 361(6400):360–365, 2018.

[3] Camille Bilodeau, Wengong Jin, Tommi Jaakkola, Regina Barzilay, and Klavs F Jensen. Generative models for molecular discovery: Recent advances and challenges. *Wiley Interdisciplinary Reviews: Computational Molecular Science*, 12(5):e1608, 2022.

[4] Juhwan Noh, Geun Ho Gu, Sungwon Kim, and Yousung Jung. Machine-enabled inverse design of inorganic solid materials: promises and challenges. *Chemical Science*, 11(19):4871–4881, 2020.

[5] Yongtae Kim, Youngsoo Kim, Charles Yang, Kundo Park, Grace X Gu, and Seunghwa Ryu. Deep learning framework for material design space exploration using active transfer learning and data augmentation. *npj Computational Materials*, 7(1):140, 2021.

[6] Claudio Zeni, Robert Pinsler, Daniel Zügner, Andrew Fowler, Matthew Horton, Xiang Fu, Sasha Shysheya, Jonathan Crabbé, Lixin Sun, Jake Smith, et al. Mattergen: a generative model for inorganic materials design. *arXiv preprint arXiv:2312.03687*, 2023.

[7] Mengjiao Yang, KwangHwan Cho, Amil Merchant, Pieter Abbeel, Dale Schuurmans, Igor Mordatch, and Ekin Dogus Cubuk. Scalable diffusion for materials generation. *arXiv preprint arXiv:2311.09235*, 2023.

[8] Tian Xie, Xiang Fu, Octavian-Eugen Ganea, Regina Barzilay, and Tommi Jaakkola. Crystal diffusion variational autoencoder for periodic material generation. *arXiv preprint arXiv:2110.06197*, 2021.

[9] W Patrick Walters and Regina Barzilay. Applications of deep learning in molecule generation and molecular property prediction. *Accounts of chemical research*, 54(2):263–270, 2020.

[10] Alexander Dunn, Qi Wang, Alex Ganose, Daniel Dopp, and Anubhav Jain. Benchmarking materials property prediction methods: the matbench test set and automatminer reference algorithm. *npj Computational Materials*, 6(1):138, 2020.

[11] Anthony Yu-Tung Wang, Steven K Kauwe, Ryan J Murdock, and Taylor D Sparks. Compositionally restricted attention-based network for materials property predictions. *Npj Computational Materials*, 7(1):77, 2021.

[12] Ya Zhuo, Aria Mansouri Tehrani, and Jakoah Brgoch. Predicting the band gaps of inorganic solids by machine learning. *The journal of physical chemistry letters*, 9(7):1668–1673, 2018.

[13] Pierre-Paul De Breuck, Geoffroy Hautier, and Gian-Marco Rignanese. Materials property prediction for limited datasets enabled by feature selection and joint learning with modnet. *npj computational materials*, 7(1):83, 2021.

[14] Logan Ward, Ankit Agrawal, Alok Choudhary, and Christopher Wolverton. A general-purpose machine learning framework for predicting properties of inorganic materials. *npj Computational Materials*, 2(1):1–7, 2016.

[15] Steven K Kauwe, Jake Graser, Ryan Murdock, and Taylor D Sparks. Can machine learning find extraordinary materials? *Computational Materials Science*, 174:109498, 2020.

[16] Zhi-Wen Zhao, Marcos Del Cueto, and Alessandro Troisi. Limitations of machine learning models when predicting compounds with completely new chemistries: possible improvements applied to the discovery of new non-fullerene acceptors. *Digital Discovery*, 1(3):266–276, 2022.

[17] Aviv Netanyahu, Abhishek Gupta, Max Simchowitz, Kaiqing Zhang, and Pulkit Agrawal. Learning to extrapolate: A transductive approach. In *International Conference on Learning Representations*, 2023.

[18] Hidetoshi Shimodaira. Improving predictive inference under covariate shift by weighting the log-likelihood function. *Journal of statistical planning and inference*, 90(2):227–244, 2000.

[19] Hiroshi Shimodaira, Masashi Sugiyama, Amos Storkey, Arthur Gretton, Shai-Ben David, J QuinoneroCandela, M Sugiyama, A Schwaighofer, and ND Lawrence. *Dataset Shift in Machine Learning*, pages 201–205. Neural Information Processing Series. Yale University Press in association with the Museum of London, 2009. ISBN 978-0-26217-005-5.

[20] Anton O Oliynyk, Erin Antono, Taylor D Sparks, Leila Ghadbeigi, Michael W Gaultois, Bryce Meredig, and Arthur Mar. High-throughput machine-learning-driven synthesis of full-heusler compounds. *Chemistry of Materials*, 28(20):7324–7331, 2016.

[21] Rdkit: Open-source cheminformatics. https://www.rdkit.org.

[22] David Weininger. Smiles, a chemical language and information system. 1. introduction to methodology and encoding rules. *Journal of chemical information and computer sciences*, 28 (1):31–36, 1988.

[23] James Damewood, Jessica Karaguesian, Jaclyn R Lunger, Aik Rui Tan, Mingrou Xie, Jiayu Peng, and Rafael Gómez-Bombarelli. Representations of materials for machine learning. *Annual Review of Materials Research*, 53(1):399–426, 2023.

[24] Gyoung S Na and Chanyoung Park. Nonlinearity encoding for extrapolation of neural networks. In *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pages 1284–1294, 2022.

[25] Eric S Muckley, James E Saal, Bryce Meredig, Christopher S Roper, and John H Martin. Interpretable models for extrapolation in scientific machine learning. *Digital Discovery*, 2(5): 1425–1435, 2023.

[26] Hajime Shimakawa, Akiko Kumada, and Masahiro Sato. Extrapolative prediction of small-data molecular property using quantum mechanics-assisted machine learning. *npj Computational Materials*, 10(1):11, 2024.

[27] Kan Hatakeyama-Sato and Kenichi Oyaizu. Generative models for extrapolation prediction in materials informatics. *ACS omega*, 6(22):14566–14574, 2021.

[28] Han Yang, Chenxi Hu, Yichi Zhou, Xixian Liu, Yu Shi, Jielan Li, Guanzhi Li, Zekun Chen, Shuizhou Chen, Claudio Zeni, et al. Mattersim: A deep learning atomistic model across elements, temperatures and pressures. *arXiv preprint arXiv:2405.04967*, 2024.

[29] Amil Merchant, Simon Batzner, Samuel S Schoenholz, Muratahan Aykol, Gowoon Cheon, and Ekin Dogus Cubuk. Scaling deep learning for materials discovery. *Nature*, 624(7990):80–85, 2023.

[30] Ilyes Batatia, Philipp Benner, Yuan Chiang, Alin M Elena, Dávid P Kovács, Janosh Riebesell, Xavier R Advincula, Mark Asta, William J Baldwin, Noam Bernstein, et al. A foundation model for atomistic materials chemistry. *arXiv preprint arXiv:2401.00096*, 2023.

[31] Kangming Li, Andre Niyongabo Rubungo, Xiangyun Lei, Daniel Persaud, Kamal Choudhary, Brian DeCost, Adji Bousso Dieng, and Jason Hattrick-Simpers. Probing out-of-distribution generalization in machine learning for materials. *arXiv preprint arXiv:2406.06489*, 2024.

[32] Stefano Curtarolo, Wahyu Setyawan, Gus LW Hart, Michal Jahnatek, Roman V Chepulskii, Richard H Taylor, Shidong Wang, Junkai Xue, Kesong Yang, Ohad Levy, et al. Aflow: An automatic framework for high-throughput materials discovery. *Computational Materials Science*, 58:218–226, 2012.

[33] Conduit G. and Bajaj S. Mechanical properties of some steels, 2017. https://citrination.com/datasets/153092/show_files/.

[34] Ioannis Petousis, David Mrdjenovich, Eric Ballouz, Miao Liu, Donald Winston, Wei Chen, Tanja Graf, Thomas D Schladt, Kristin A Persson, and Fritz B Prinz. High-throughput screening of inorganic compounds for the discovery of novel dielectric and optical materials. *Scientific data*, 4(1):1–12, 2017.

[35] Anubhav Jain, Shyue Ping Ong, Geoffroy Hautier, Wei Chen, William Davidson Richards, Stephen Dacek, Shreyas Cholia, Dan Gunter, David Skinner, Gerbrand Ceder, et al. Commentary: The materials project: A materials genome approach to accelerating materials innovation. *APL materials*, 1(1), 2013.

[36] Bharath Ramsundar, Peter Eastman, Patrick Walters, Vijay Pande, Karl Leswing, and Zhenqin Wu. *Deep Learning for the Life Sciences*. O'Reilly Media, 2019.

[37] Esther Heid, Kevin P Greenman, Yunsie Chung, Shih-Cheng Li, David E Graff, Florence H Vermeire, Haoyang Wu, William H Green, and Charles J McGill. Chemprop: a machine learning package for chemical property prediction. *Journal of Chemical Information and Modeling*, 64 (1):9–17, 2023.

[38] Leo Breiman. Random forests. *Machine learning*, 45:5–32, 2001.

[39] Matt W Gardner and SR Dorling. Artificial neural networks (the multilayer perceptron)—a review of applications in the atmospheric sciences. *Atmospheric environment*, 32(14-15): 2627–2636, 1998.

[40] Logan Ward, Alexander Dunn, Alireza Faghaninia, Nils ER Zimmermann, Saurabh Bajaj, Qi Wang, Joseph Montoya, Jiming Chen, Kyle Bystrom, Maxwell Dylla, et al. Matminer: An open source toolkit for materials data mining. *Computational Materials Science*, 152:60–69, 2018.

[41] Alexander Kraskov, Harald Stögbauer, and Peter Grassberger. Estimating mutual information. *Physical Review E—Statistical, Nonlinear, and Soft Matter Physics*, 69(6):066138, 2004.

A Appendix

A.1 Additional Results

In Table 2, we compare the TPR of OOS detection, showcasing our ability to produce candidate materials with outstanding property values, with an improvement of 3x compared with the strongest baseline. In Table 3 we demonstrate that all methods perform well, making Bilinear Transduction a strong predictor both in and out of support alike. In Figure 4 we show extrapolation for a molecular property, Freesolv.

Table 2: **Solids (top) and molecules (bottom)** OOS True Positive Rate (TPR).

Dataset	Property	#Samples	Ridge Reg. [15]	MODNet [13]	CrabNet [11]	Ours
AFLOW [32]	Band Gap [eV]	14123	0.0	0.0	0.032	0.132
	Bulk Modulus [GPa]	2740	0.124	0.007	0.199	0.336
	Debye Temperature [K]	2740	0.182	0.0	0.0	0.504
	Shear Modulus [GPa]	2740	0.058	0.088	0.044	0.182
	Thermal Conductivity [$\frac{W}{mK}$]	2734	0.088	0.0	0.0	0.132
	Thermal Expansion [K^{-1}]	2733	0.154	0.154	0.132	0.191
Matbench [10]	Band Gap [eV]	2154	0.0	0.004	0.0	0.009
	Refractive Index	4764	0.009	0.101	0.004	0.122
	Yield Strength [MPa]	312	0.0	0.0	0.0	0.0
MP [35]	Bulk Modulus [GPa]	6307	0.073	0.003	0.311	0.498
	Elastic Anisotropy	6331	0.0	0.0	0.0	0.006
	Shear Modulus [GPa]	6184	0.0	0.0	0.003	0.084
			Chemprop [37]	**Random Forest** [38]	**MLP** [39]	
MoleculeNet [36]	ESOL [$\frac{mol}{L}$]	1128	0.357	0.0	0.196	0.268
	Freesolv [$\frac{kJ}{mol}$]	643	0.062	0.0	0.0	0.781
	Lipophilicity [$\log D$]	4200	0.024	0.0	0.014	0.057
	BACE binding [IC50]	1513	0.0	0.0	0.0	0.013

Table 3: **Solids (top) and molecules (bottom)** In-distribution mean average prediction error and standard error of the mean.

Dataset	Property	Ridge Reg. [15]	MODNet [13]	CrabNet [11]	Ours
AFLOW [32]	Band Gap [eV]	0.87 ± 0.04	0.56 ± 0.02	0.35 ± 0.02	0.61 ± 0.02
	Bulk Modulus [GPa]	15.41 ± 1.21	15.1 ± 1.3	8.01 ± 1.05	13.06 ± 1.6
	Debye Temperature [K]	0.13 ± 0.01	0.13 ± 0.01	0.09 ± 0.01	0.14 ± 0.01
	Shear Modulus [GPa]	0.31 ± 0.03	0.27 ± 0.02	0.19 ± 0.02	0.31 ± 0.03
	Thermal Conductivity [$\frac{W}{mK}$]	0.47 ± 0.03	0.4 ± 0.02	0.31 ± 0.03	0.43 ± 0.04
	Thermal Expansion [K^{-1}]	0.11 ± 0.01	0.07 ± 0.01	0.04 ± 0.0	0.11 ± 0.01
Matbench [10]	Band Gap [eV]	1.75 ± 0.07	0.32 ± 0.03	0.24 ± 0.03	0.49 ± 0.05
	Refractive Index	1.00 ± 0.05	0.15 ± 0.01	0.13 ± 0.02	0.16 ± 0.01
	Yield Strength [MPa]	411 ± 75	62.5 ± 11.8	52.4 ± 18.1	156 ± 33
MP [35]	Bulk Modulus [GPa]	36.9 ± 1.21	18.63 ± 1.16	10.2 ± 0.8	19.4 ± 1.3
	Elastic Anisotropy	22.00 ± 2.01	2.12 ± 0.34	1.24 ± 0.06	2.4 ± 0.3
	Shear Modulus [GPa]	35.7 ± 1.2	12.8 ± 0.7	8.75 ± 0.63	13.6 ± 0.7
		Chemprop [37]	**Random Forest** [38]	**MLP** [39]	
MoleculeNet [36]	ESOL [$\frac{mol}{L}$]	0.28 ± 0.03	0.25 ± 0.03	0.28 ± 0.03	0.29 ± 0.04
	Freesolv [$\frac{kJ}{mol}$]	0.16 ± 0.02	0.20 ± 0.06	0.18 ± 0.06	0.12 ± 0.02
	Lipophilicity [$\log D$]	0.36 ± 0.02	0.40 ± 0.02	0.38 ± 0.03	0.46 ± 0.03
	BACE binding [IC50]	0.45 ± 0.04	0.37 ± 0.04	0.43 ± 0.05	0.51 ± 0.05

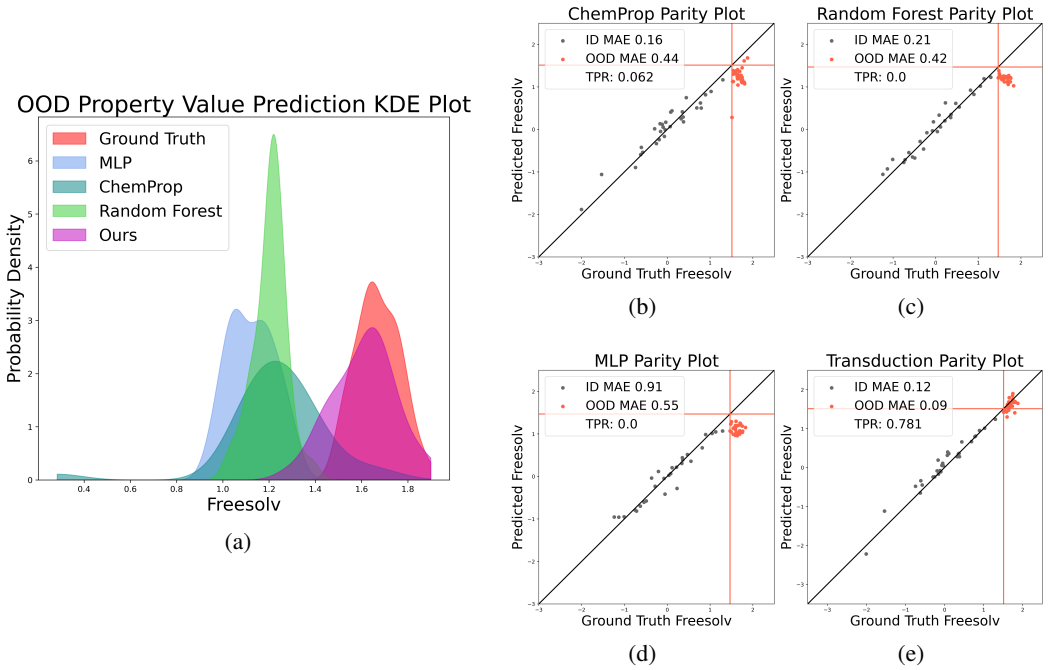


Figure 4: **Left: OOS prediction.** Our transductive approach predicts values closer to the desired distribution (a). **Right: In-distribution, and OOS Freesolv predictions vs. ground truth values**. While (b) Chemprop [37], (c) RF, (d) MLP and (e) Bilinear Transduction (ours) perform well within the training distribution (gray dots bounded by the red horizontal line), only Bilinear Transduction performs well beyond this range on OOS data (red dots).

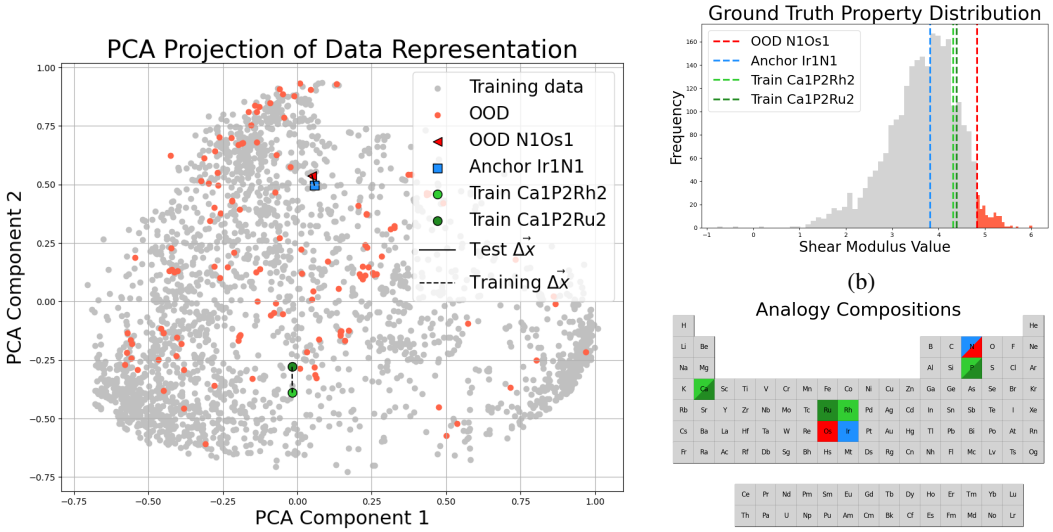
A.2 Additional Results For Analogy Analysis

A.2.1 Solids

Figure 5 describes shear modulus inference of OOS NOs via anchor NIr, analogous to training anchor CaP_2Rh_2 and training target CaP_2Ru_2. In this case, the training anchor and target (CaP_2Rh_2 and CaP_2Ru_2) and test anchor and target (NIr and NOs) differ by one d-block element: Rh (Z=45) to Ru (Z=44) and Ir (Z=77) to Os (Z=76). Figure 6 describes bulk modulus inference of OOS N_3Nb_4 via anchor HfNbP, analogous to training anchor HfMoP and training target Mo_8P_5. Additional examples include shear modulus prediction for OOS NbSiIr via anchor NbSiPt analogous to training anchor GeNbIr and target GeNbPt, and OOS ReSiNb via anchor OsSiZr analogous to training anchor GeIrNb and target GeIrPt.

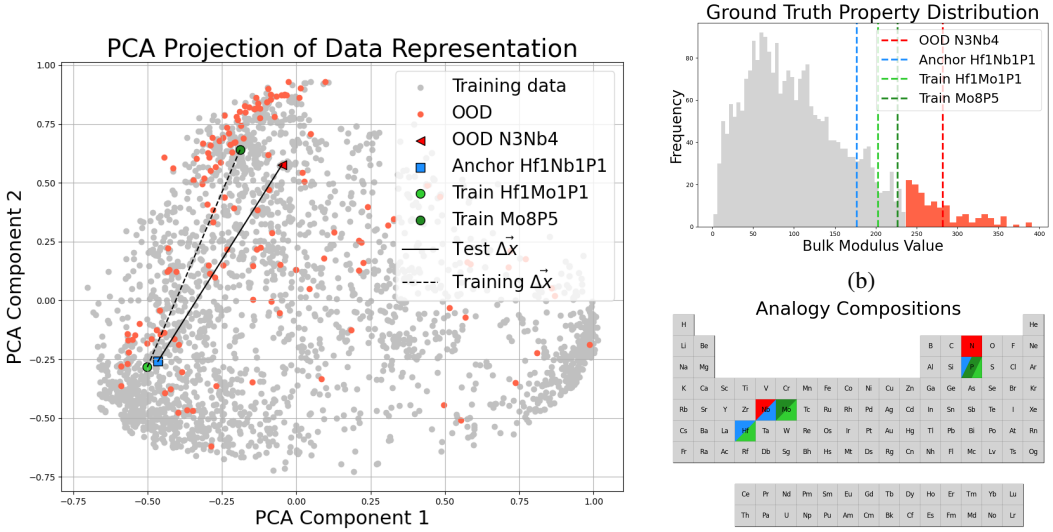
A.2.2 Molecules

Figures 7 and 8 display additional analogical molecule pairs. Two distinct modes of similarity can be identified: one is between the training anchor and target, and between the test anchor and target Figures 8a, 8b, 8c, and the other is between the anchors, and between the targets Figures 8d, 8e, 8f. In the anchor selection process, the model can converge to an anchor that is either very similar to the OOS target, in a way that two training samples are similar (in $\mathcal{X}$ space), or that is different from the OOS target and similar to the training anchor and the training target will be similar to the OOS. For both modes, we can spot analogous differences between the molecules.



(a) (b) (c)

Figure 5: **Visualizing Analogies in Bilinear Transduction**. AFLOW shear modulus OOS predictions are based on in-distribution anchors, that paired with OOS points, form analogies to training pairs. (a) PCA plot of all samples in the dataset. The difference between the OOS point and its anchor is similar to the difference between the training point and anchor. (b) Ground truth shear modulus training (gray) and test (red) distributions and OOS, anchor, and analogous training pair values. (c) Analogy compositional visualization. anchor and OOS differ by one neighboring d-block element. So do training anchor and training target.



(a) (b) (c)

Figure 6: **Visualizing Analogies in Bilinear Transduction**. AFLOW bulk modulus OOS predictions are based on in-distribution anchors, that paired with OOS points, form analogies to training pairs. (a) PCA plot of all samples in the dataset. The difference between the OOS point and its anchor is similar to the difference between the training point and anchor. (b) Ground truth shear modulus training (gray) and test (red) distributions and OOS, anchor, and analogous training pair values. (c) Analogy compositional visualization.

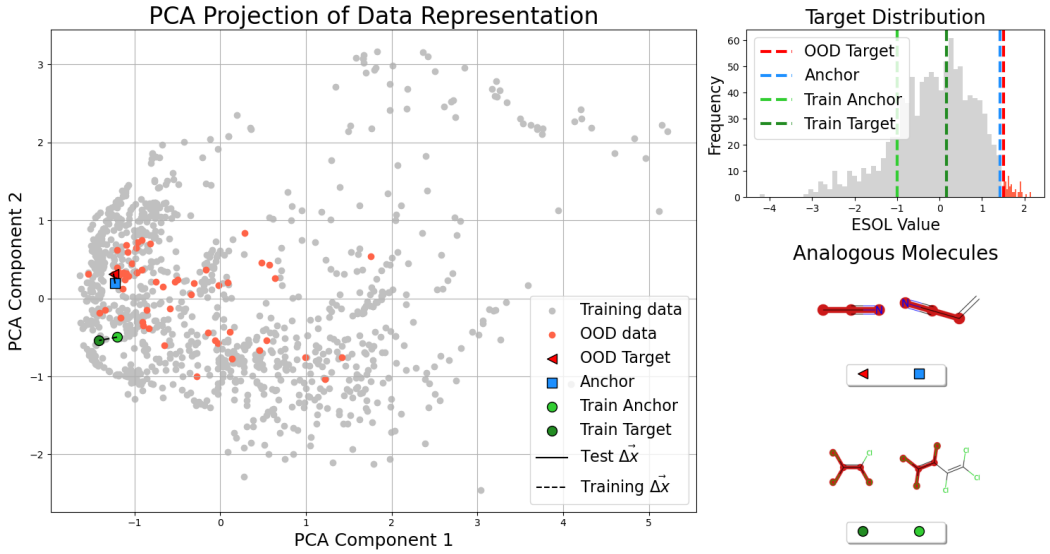


Figure 7: **Bilinear Transduction Analogy Visualization**. MoleculeNet ESOL OOS predictions are based on differences between in-distribution anchors and OOS targets, that form analogies to training pairs. (a) OOS-anchor and training pair differences. (b) Ground truth training (gray) and test (red) distributions and OOS, anchor, and analogous training pair values. (c) Analogous molecule pairs. The similarity between OOS-anchor and training anchor and target is highlighted in red, using the MCS metric.

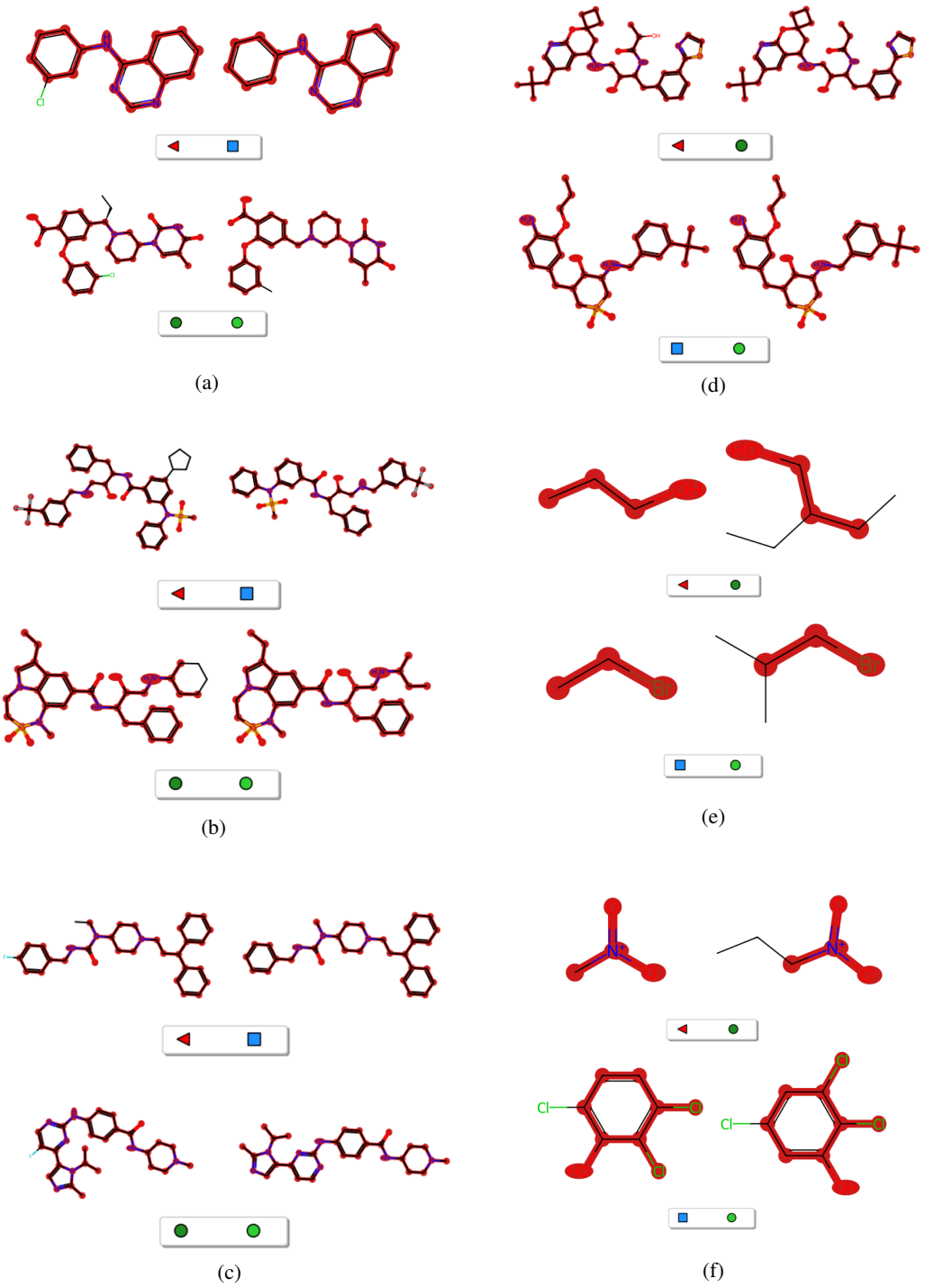


Figure 8: **Analogical Molecules pairs.** Molecules are paired according to maximum similarity. (a-c) **top** pair: OOS are similar to their anchors, **bottom** pair: training targets are similar to their anchors. (d-f) **top** pair: OOS are similar to training targets, **bottom** pair: anchors are similar to training anchors. For each pair, we denote the relevant benchmark and chemical operator differentiating samples within each pair. **Left** Columns: (a) lipophilicity, Cl addition. (b) BACE, addition or completion of a ring. (c) lipophilicity, F addition. (d) BACE, targets differ in OH functional group and anchors are the same. (e) ESOL, additional C. (f) ESOL, targets differ in additional C and anchors differ in functional group position in the ring.

A.3 Implementation Details

A.3.1 Data Representation

Ridge Regression. We adhere to the data preprocessing scheme outlined in Kauwe et al. [15] on AFLOW and featurize data using element-based Oliynyk descriptors [20]. The data is scaled using the StandardScaler and normalized using the Normalizer from the sklearn library based on the training data statistics.

MODNet. Following De Breuck et al. [13] on Matbench, we create element-based feature vectors using Matminer [40], and then select features based on the normalized mutual information [41]. The data is scaled using the MinMaxScaler based on the training data statistics.

CrabNet. Following Wang et al. [11] on MP, we leverage mat2vec, learned via self-supervised natural language processing techniques, trained on a large corpus of scientific literature.

Bilinear Transduction. We use the Ridge Regression representation and processing for AFLOW, and the MODNet representation and processing for Matbench. For MP, we use a descriptor-based representation, Oliynyk [20], scaled with MinMaxScaler from the sklearn library based on the training data statistics.

A.3.2 Bilinear Transduction Hyperparameter Search

In Table 1, we report the best OOS MAE score for Bilinear Transduction on AFLOW over a hyperparameter search on the number of predictor network layers $(3, 4)$, layer size $(256, 512, 1024)$ and embedding size $(42, 48, 64)$. The hyperparameter search revealed little sensitivity to changes in hyperparameter values, indicating the robustness of our evaluation.