

Substrate Scope Contrastive Learning: Repurposing Human Bias to Learn Atomic Representations

Wenhao Gao,[†] Priyanka Raghavan,[†] Ron Shprints,[†] and Connor W. Coley^{*,†,‡}

[†]*Department of Chemical Engineering, MIT, Cambridge, MA*

[‡]*Department of Electrical Engineering and Computer Science, MIT, Cambridge, MA*

E-mail: ccoleymit.edu

Abstract

Learning molecular representation is a critical step in molecular machine learning that significantly influences modeling success, particularly in data-scarce situations. The concept of broadly *pre-training* neural networks has advanced fields such as computer vision, natural language processing, and protein engineering. However, similar approaches for small organic molecules have not achieved comparable success. In this work, we introduce a novel pre-training strategy, *substrate scope contrastive learning*, which learns atomic representations tailored to chemical reactivity. This method considers the grouping of substrates and their yields in published substrate scope tables as a measure of their similarity or dissimilarity in terms of chemical reactivity. We focus on 20,798 aryl halides in the CAS Content Collection™ spanning thousands of publications to learn a representation of aryl halide reactivity. We validate our pre-training approach through both intuitive visualizations and comparisons to traditional reactivity descriptors and physical organic chemistry principles. The versatility of these embeddings is further evidenced in their application to yield prediction, regioselectivity

prediction, and the diverse selection of new substrates. This work not only presents a chemistry-tailored neural network pre-training strategy to learn reactivity-aligned atomic representations, but also marks a first-of-its-kind approach to benefit from the human bias in substrate scope design.

Introduction

Encoding molecular structures into appropriate numerical representations that are computer-readable is key to accurate prediction of molecules' reaction behavior, which is a pivotal challenge in chemical science and engineering. Traditionally, explaining reactivity based on chemical structure has relied on physical modeling and mechanistic analysis, along with the computation of key physical descriptors to build statistical models.^{1–4} This “feature engineering” approach relies on a prior understanding of relevant (computable) properties, which vary across different classes of molecules and reactions. In contrast, the advent of machine learning (ML)⁵ and, notably, deep learning (DL),⁶ has facilitated end-to-end learning as an alternative to manual crafting of molecular features. In principle, DL models can extract task-specific information from chemical structures, forming numerical features known as *representations* or *embeddings*. While already applied in modeling reaction yields,^{7–9} most DL models remain remarkably ineffective or inaccurate in low data regimes.

The paradigm of pre-training deep neural networks on tasks with abundant data and adapting them to other tasks^{10–14} offers an attractive approach to enhance efficiency and generalizability. Ideally, such strategies enable few-shot learning, where models can learn new concepts from just a handful of examples.^{15,16} Successes in this approach are evident across computer vision,^{17–20} natural language processing,^{21–23} and protein engineering.^{24,25} Inspired by these successes, numerous network pre-training strategies for molecular structures have been developed.^{26–30} However, unlike in other domains, these methods have not yet led to substantial improvements in modeling.³¹

We believe that there is a fundamental misalignment between existing pre-training tasks

and the information crucial for downstream applications. Pre-training tasks for graph neural networks, for example, have involved predicting atom or functional group identities given the rest of the chemical structure,^{26,32} which primarily teaches models about valence rules and the prevalence of certain functional groups. Similarly, pre-training tasks for models operating on SMILES strings³³ teach models about similar principles along with the specifics of the SMILES language.^{34–37} Recently introduced pre-training tasks involving the reconstruction of three-dimensional (3D) conformations, add steric considerations to the spectrum of information a model can learn,^{28,30,38} which represents progress but is still coarse for many of the chemical properties we are interested in.

Anticipating chemical reactivity³⁹—whether quantitatively (e.g., predicting yields) or qualitatively (e.g., predicting regioselectivity)—is a challenge where an ideal pre-trained molecular representation would incorporate information beyond valence rules, substructure popularity, and accessible conformations. To address this challenge, we introduce **substrate scope contrastive learning** (ContraScope), a new pre-training approach based on contrastive learning^{40–42} that leverages published reactions and benefits from human bias in substrate scope design (see Figure 1B and more statistics in Figure S2). The core idea of our method is that molecules exhibiting similar reactivity under identical circumstances should be considered similar in numerical representation as well. Given that one molecule can react in many disparate ways, we focus our representation learning goals on obtaining atom-level embeddings of specific reactive sites, rather than molecule-level embeddings. As a measure of synthetic similarity, we utilize substrate scope tables—collections of substrates, typically numbering 10-30, with their reaction yields reported under a consistent set of conditions. Published substrate scope tables often reflect a bias towards successful outcomes.⁴³ However, this bias suggests that, *on average*, two molecules within the same scope exhibit more similar reactivity than two molecules from different scopes. ContraScope is devised based on this principle.

We demonstrate the efficacy of our contrastive learning approach using aryl halides, a

key class of substrates employed in a wide range of chemical reactions. Specifically, we curated substrate scope tables from the literature published from 2010 to 2015, as indexed in the CAS Content CollectionTM, the largest human-curated collection of scientific data in the world, and demonstrate that our pre-training strategy successfully learns representations that capture key aspects of aryl halide reactivity. The learned embeddings show quantitative correlations with traditional reactivity descriptors and can be directly applied to tasks with limited data, such as yield prediction, regioselectivity prediction, and substrate selection. Our results support the hypothesis that substrate scope tables and publication bias, though previously considered detrimental to model training, actually offer valuable insights into reactivity patterns and can serve as a significant source of information. We hope to encourage the development of additional functionality-focused pre-training tasks—and pre-training tasks that effectively utilize human bias—to further the pursuit of universal molecular representations in chemistry.

Substrate Scope Contrastive Learning

The goal of our pre-training approach is to derive a multidimensional vector representation for the reactive atom in an aryl halide substrate that captures essential aspects of its reactivity. Ideally, this could be accomplished through supervised learning using a vast and varied dataset of molecules and their respective yields across numerous reaction conditions. However, the reality of data acquisition presents us with substrate scope tables that represent a narrow selection of molecules, influenced by the selective reporting tendencies and norms in scientific research and publication (see Figure S2 for statistics of the dataset). The core idea of substrate scope contrastive learning (ContraScope) is to use molecules not listed in a scope table as negative samples, leveraging the artificial bias in the substrate scope selection.

The model is trained on triplet of molecules: for each molecule as an anchor, we randomly select another molecule from the same substrate scope to serve as a positive sample, ensuring

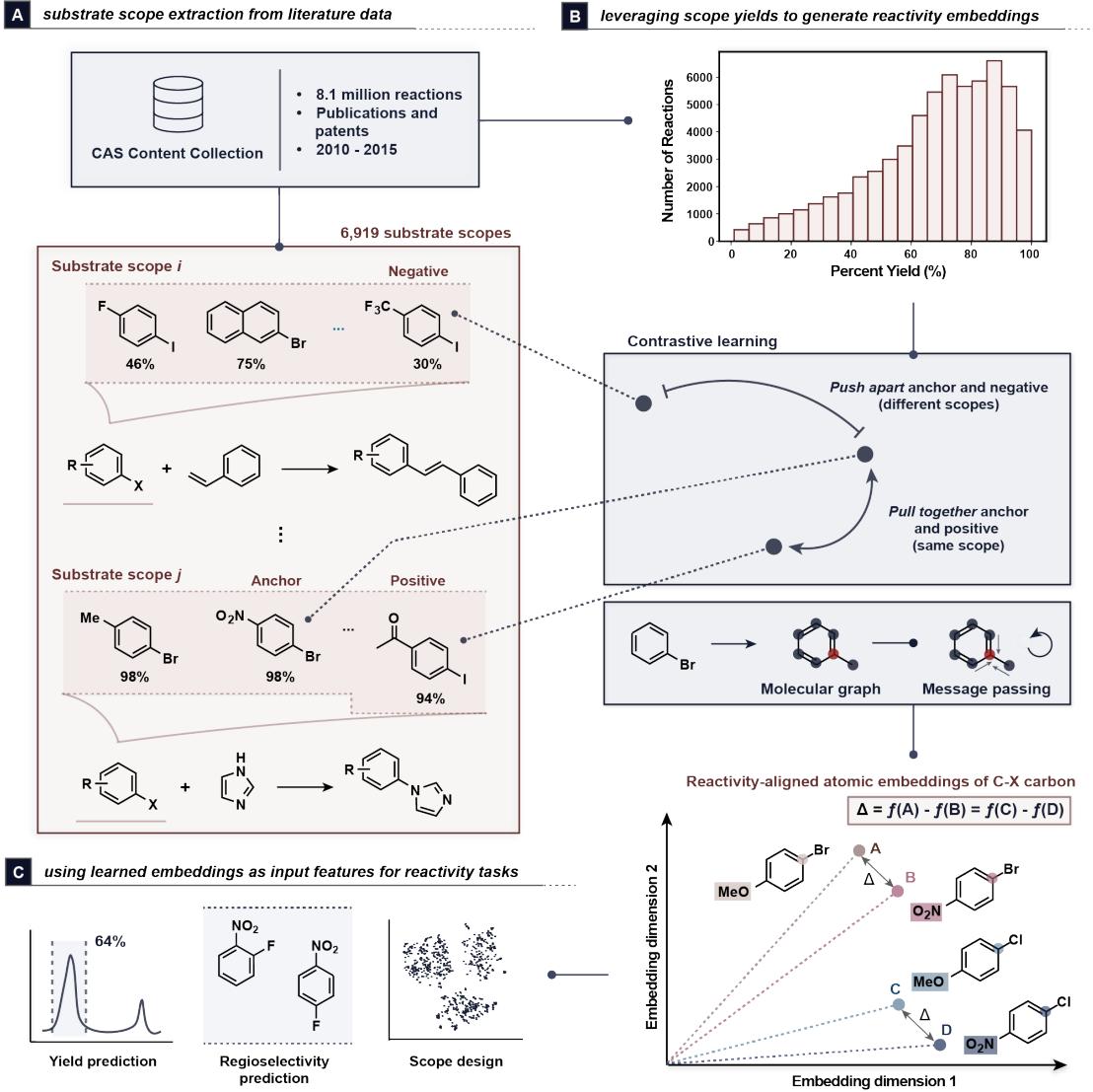


Figure 1: (A) Substrate scope tables for training the network are curated from the CAS Content Collection, focusing on aryl halides with their associated yields. The substrate scope *i*, *j* shown in the figure are real samples from the database.^{44,45} (B) Overview of *substrate scope contrastive learning*: Our pre-training strategy leverages the selective reporting bias in chemical reactions, partially revealed by the pronounced inclination towards higher yields in the histogram displaying the yield distribution for all surveyed reactions. Embeddings of substrates from the same scope table are pulled together, while embeddings of substrates from distinct scope tables are forced apart. A message-passing neural network operates on molecular graphs to derive atomic embeddings. The training aims to provide reactivity-aligned atomic embeddings learned from the substrate scopes, constructing a vector space representing reactivity trends. (C) The uniformly pre-trained embeddings can benefit various reactivity-related tasks.

that their embeddings are closely aligned; concurrently, we sample a molecule that doesn’t belong to the substrate scope to act as a negative sample, ensuring their embeddings are distantly separated. Recognizing that molecules within the same substrate scope can exhibit varying degrees of reactivity, we define the ideal distance between the anchor and positive molecules as the difference in their yields. Formally, we train our model by minimizing the following loss function:

$$\mathcal{L}(m_a, m_p, m_n) = \underbrace{(d(m_a, m_p) - \gamma|y_a - y_p|)^2}_{\text{Distance proportional to yield difference}} + \underbrace{\beta \log[1 + \exp(M - d(m_a, m_n))]}_{\text{Distance at least a margin M away}} \quad (1)$$

where m_a, m_p, m_n denote a triplet of molecules, comprising an anchor, a positive, and a negative sample. $d(\cdot, \cdot)$ denotes the embedding distance between two molecules as learned by the model. y_a and y_p represent the reported yield under identical reaction conditions for the anchor and positive molecules. The margin M , alongside the ratios β and γ are constants determined through hyperparameter tuning (see section *Learning curves and hyper-parameter tuning* in SI for detailed values). These terms “pull” substrates from the same substrate scope closer based on how similar their yields are, while they “push” away all other aryl halides, on average, not reported in the scope.

Substrate scope tables and associated reaction yields are sourced from a subset of the CAS Content Collection describing publications and patents between 2010 and 2015. We restrict our analysis to aryl halides as they are an important category of molecules widely employed as building blocks in medicinal chemistry.^{46,47} To ensure consistency within each substrate scope, we verified that all reactions involved transformations at a C–X bond and were performed under identical conditions, including reactants besides the aryl halides in the case of multi-component reactions, according to the database. We excluded scopes with fewer than five reactions, resulting in a total of 6,919 scopes with 64,192 reactions. For each training epoch, each aryl halide serves as the anchor in 16 triplets, which also include one randomly-sampled positive instance from the same scope and one randomly-

sampled negative instance from a different scope. We adopt a graph isomorphism network (GIN),⁴⁸ a widely used graph neural network, as the model architecture. The aryl halides are represented in graphs with atom and bond identities as input features.⁴⁹ Details on our training methodology are in the Methods Section.

Results

Learned aryl halide embeddings exhibit alignment with qualitative and quantitative measures of reactivity

To evaluate the consistency of the learned aryl halide representations with human understanding of reactivity, we perform a combination of qualitative and quantitative analyses. We first visualize the learned embeddings to elucidate how the model has learned to organize the chemical space of aryl halides. To this end, we encoded a set of molecules comprising the 500 most frequently used aryl halides, augmented by a random sample of 2922 aryl halides from the substrate scope dataset. Upon encoding, we used t-distributed stochastic neighbor embedding (t-SNE)⁵⁰ projection to visualize the 64-dimensional embedding learned by the model in Figure 2A (see Figure S8 for results of other projection methods and a comparison with embeddings from an untrained network). A detailed examination of the position of specific structures in the embedding space (referred to as call-outs in Figure 2A) reveals that molecules with qualitatively similar reactivities—characterized by either electron-withdrawing groups (e.g., nitro, aldehyde, carbonyl) or electron-donating groups (e.g., hydroxyl, ether, amine, alkyl)—tend to cluster together. This pattern is consistent across various halide classes, underscoring the broad applicability of the embeddings we have developed.

In order to validate whether quantitative distances in the learned embedding space accurately reflect known functional similarity in chemical reactivity, we next encoded a set of 12 hand-selected para-substituted bromobenzenes, then calculated and visualized their pair-

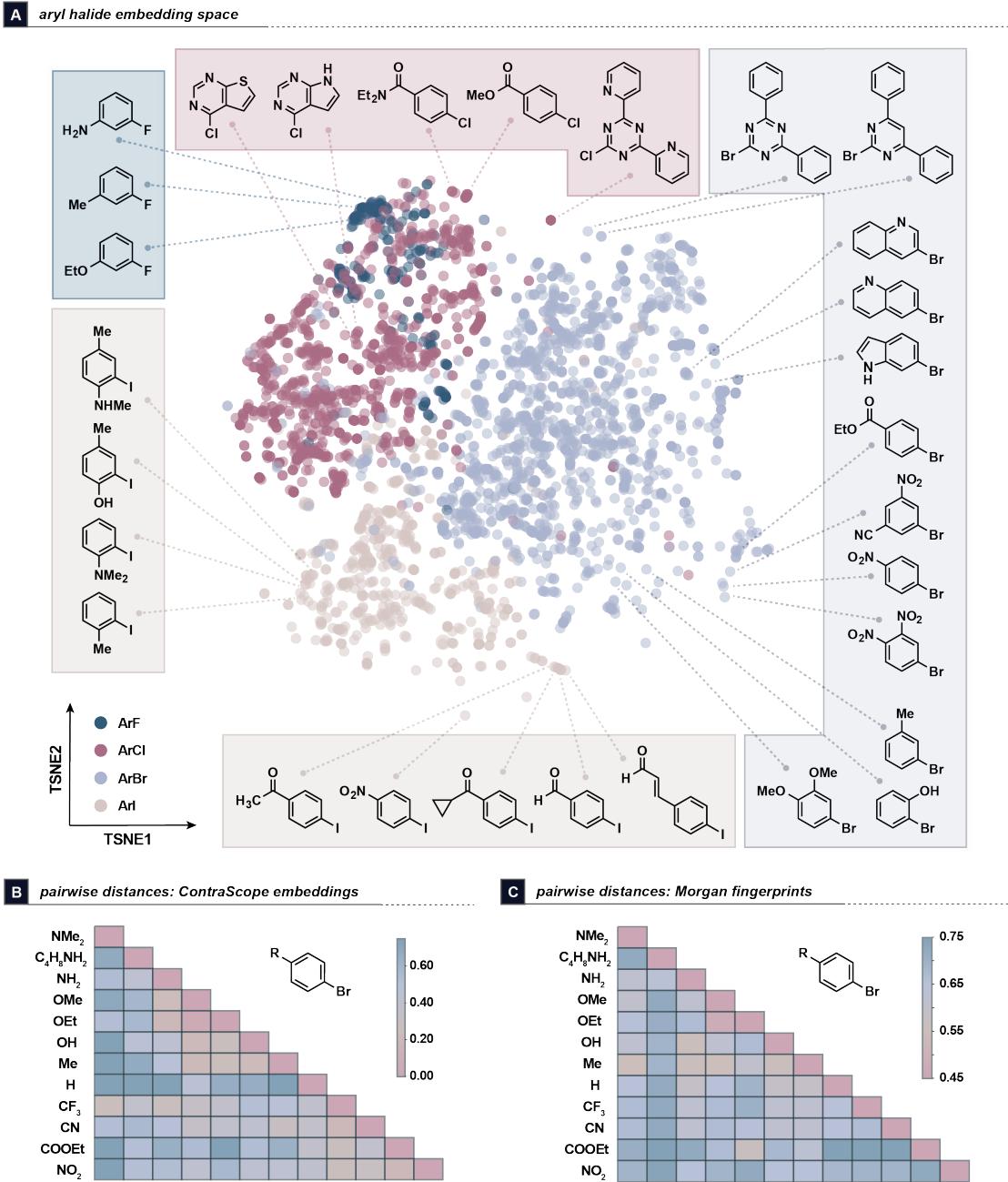


Figure 2: Qualitative analysis of aryl halide embeddings obtained through substrate scope contrastive learning. (A) A two-dimensional projection of learned embeddings using t-SNE,⁵⁰ each point denotes an aryl halide and is colored by halide type. Neighborhoods of similar aryl halides in the embedding space are annotated with structures, with the relevant C–X carbon highlighted. (B–C) Heatmap of the pair-wise distance between 12 para-substituted aryl bromides, showcasing that the learned embeddings' signal-to-noise (SNR) distances (left) better reflects functional similarity than the conventional Tanimoto distances based on structural fingerprints (right).

wise distances (Figure 2B). The heatmap reveals that molecules with electron-withdrawing groups, such as nitrile, ester, and trifluoromethyl, demonstrate notably reduced distances compared to those with electron-donating groups like methoxy, hydroxyl, ether, and amine. This discernible pattern is not mirrored in the Tanimoto distances^{51,52} (refer to Figure 2C), where distances are measured based on shared structural features in Morgan fingerprints. This comparison emphasizes that distances in our learned embedding space more effectively capture functional similarities in chemical reactivity than standard measures of structural similarity.

To further probe the correlation between our data-driven, substrate scope-informed embeddings and chemical reactivity, we embarked on a comparative analysis employing established reactivity descriptors² calculated with first-principle quantum calculations.⁵⁴ We analyze a set of 762 aryl halides comprising of 500 of the most frequently appearing aryl halides from our dataset and 262 monosubstituted aryl halides with functional groups sourced from a Hammett constant table.⁵⁵ We calculated reactivity descriptors using Gaussian 16⁵⁶ and auto-qchem,⁵⁷ with a particular focus on local atomic properties at the aromatic carbon bonded to the halogen, as our contrastive learning approach is designed to learn reactivity localized to the C–X motif, rather than molecule-level properties. Subsequent t-SNE projection mapping of this aryl halide set, as illustrated in Figure 3A, revealed that molecules positioned closely in the embedding space exhibited similar values of these atom-level reactivity descriptors. While the strongest differentiator is the identity of the halogen, more subtle patterns within each halide class can be seen.

To provide a more quantitative perspective of consistency, we trained simple (linear) machine learning models to predict various reactivity descriptors using the ContraScope embeddings as input features. We compare model performance using our embeddings, widely-used molecular descriptors⁵⁸ or fingerprints,⁵⁹ and embeddings from two other popular pre-trained deep learning models. One is the same GIN architecture but pre-trained with a common attribute masking task (AttrMask),²⁶ and another is a state-of-the-art chemical language model

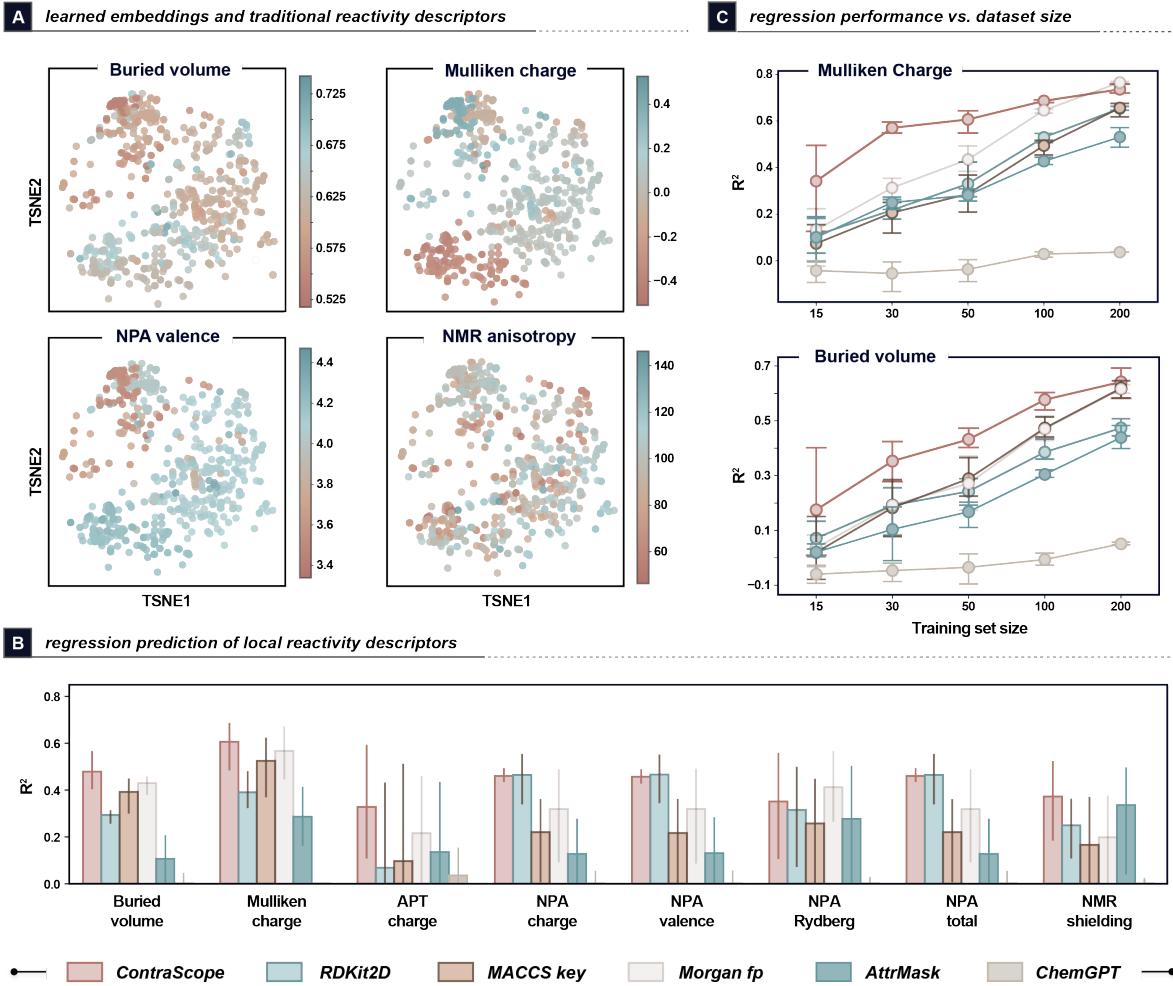


Figure 3: Relationship between the learned embeddings and conventional reactivity descriptors. (A) t-SNE visualizations of the learned embeddings, colored by traditional reactivity indicator values. (B) Regression performance (r^2) when learning to predict various physical organic chemistry descriptors from different input representations using linear models; negative values are excluded from the plot. Our embeddings, ContraScope, exhibit enhanced correlation with calculated reactivity descriptors compared to other common features. (C) Analysis of support vector machine (SVM)⁵³ regression performance as a function of dataset size, highlighting the efficacy of our embeddings in scenarios with limited data.

(ChemGPT).³⁷ The regression performance of embeddings learned through substrate scope contrastive learning was comparable to, and often exceeded, these other representations (3B). As expected, these correlations are weaker for molecule-level features (Figure S16). Moreover, we also observe a stronger ability to predict reactivity descriptors starting from our embeddings in data-limited regimes (Figure 3C), highlighting the effectiveness of our

pre-training strategy.

While these descriptors are all computable and do not truly need to be predicted from other molecular representations, these results illustrate a proof-of-concept for fine-tuning on reactivity-related tasks. The success of our purely data-driven approach, rooted in substrate scope selection, underscores the overlooked value of the human bias inherent in substrate scope tables, which unlocks a new information source for chemical reactivity modeling.

Pre-trained embeddings help enable various downstream applications

The primary motivation for pursuing better molecular representations through pre-training is to enable downstream applications, particularly in low data regimes. We therefore select three example use cases with which to illustrate the promise of substrate scope contrastive learning.

Yield prediction. As a first case study, we examine the prediction of reaction yields of various aryl bromides under identical reaction conditions. We select a Ni/photoredox catalyzed cross-coupling reaction, with reported yields for a diverse set of substrates published after 2015,³ and therefore not contained in our pre-training set. Kariofillis et al. report achieving a validation r^2 of 0.57 with a univariate model based on a computed DFT descriptor, specifically electronegativity. Without being trained on any physical organic chemistry concepts or features (like electronegativity) and without requiring the cost of electronic structure calculations, our learned ContraScope embedding is able to achieve a comparable r^2 of 0.51 (Figure 4A). Other common fixed or pre-trained representations based on structure are far less successful.

Regioselectivity prediction. As a second case study, we applied our embeddings to compare likelihood of multiple reactive sites within a molecule. We focused on a palladium-catalyzed Suzuki-Miyaura coupling reaction using polyfluoronitrobenzenes.⁶⁰ The model was trained with yield data from penta- and trifluoronitrobenzenes, and used to predict the reac-

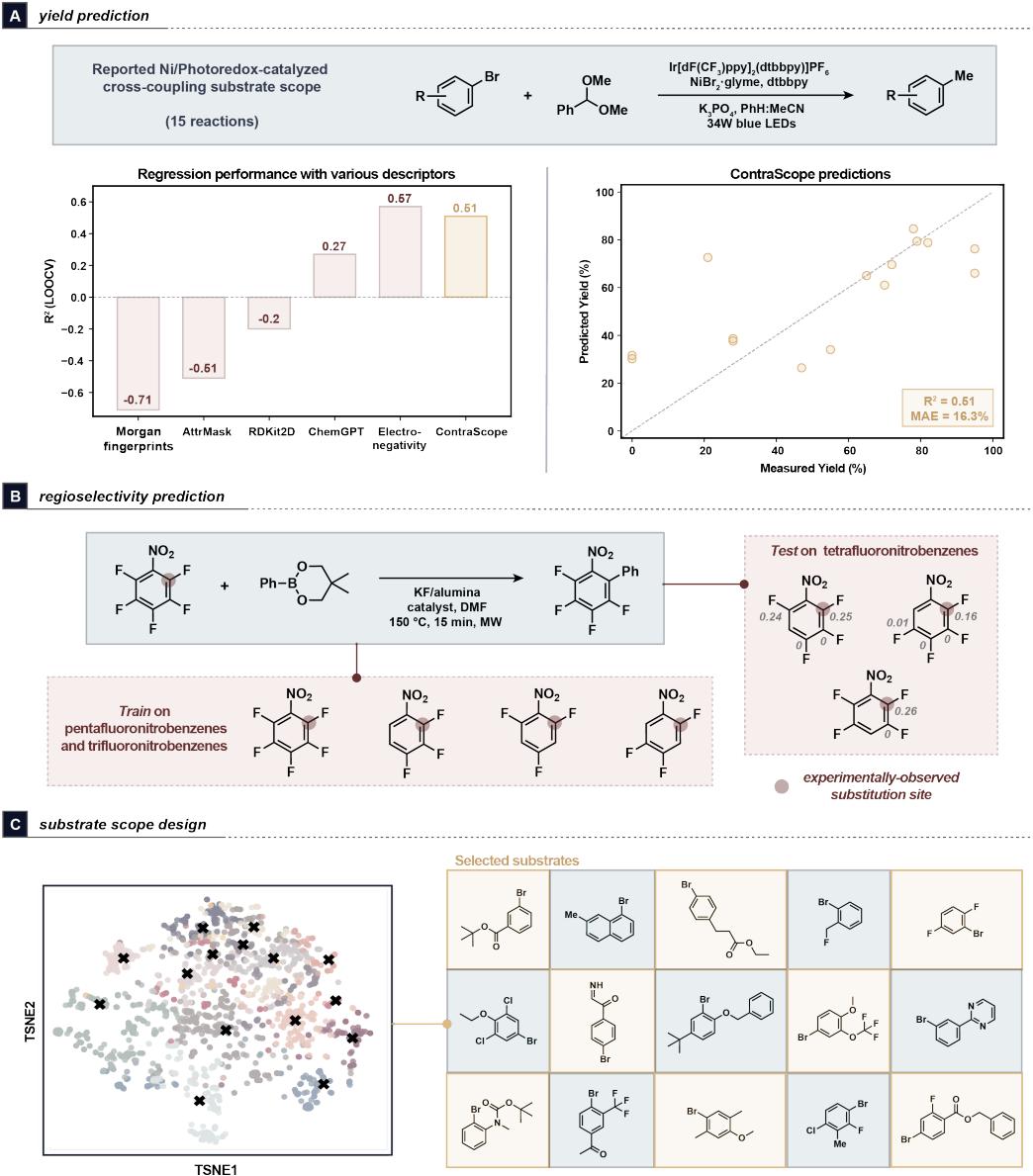


Figure 4: Validation and application of learned embeddings in various downstream tasks. (A) Application of our embeddings to predict the yield of aryl bromides in cross-coupling reactions.³ We show its predictive performance via leave-one-out validation compared with other common featurizations and a scatter plot mapping our predictions against experimental yields. (B) Application of our embeddings to predict the regioselectivity in arylation reactions of fluorobenzenes, with the model trained on penta- and trifluoronitrobenzenes to anticipate reactivity in tetrafluoronitrobenzenes, highlighting the reaction centers confirmed by experiments and annotated with prediction outcomes.⁶⁰ (C) Application of our learned embeddings for substrate scope selection, illustrated by a t-SNE plot representing purchasable aryl bromides categorized by clustering, with the chosen scope marked by cross symbols and detailed on the side.

tivity of different C–F sites in tetrafluoronitrobenzenes. The results, qualitatively confirmed by experimental data, show that a support vector machine (SVM) model using our aryl halide representations correctly learns to predict the expected reactivity at the C–F site *ortho* to the nitro group, but also distinguishes subtle differences in electronic effects, particularly where two adjacent C–F sites are nonequivalent (Figure 4B). This result highlights the proficiency of our embeddings in providing atomic-level representations to distinguish similar reactive sites and underscores the importance of learning atomic embeddings rather than molecular embeddings for reactivity prediction tasks.

Substrate scope selection. As our final case study, we illustrate how learned embeddings can assist in the selection of a “diverse” set of aryl bromides, which one may wish to do at the outset of an experimental screening campaign. This process, inspired by Kutchukian et al.’s⁶¹ and Kariofillis et al.’s³ methodology, involved clustering the chemical space of commercially-available aryl bromides using K-means;⁶² while the authors used a DFT-derived feature vector, here we use our learned embeddings. A representative aryl bromide is chosen from each cluster to form a set designed to exhibit a diverse range of reactivities (Figure 4C). Unlike clustering based on structural fingerprints or expert-selected descriptors, our approach relies on the specific reactivity profiles of aryl halides as exemplified by our embeddings, in principle leading to a more curated set for this class of compounds better aligned with their unique reactivity characteristics. As there is no objective correct answer for diverse substrate selection, we leave the selection for qualitative interpretation by the reader.

Discussion

We have demonstrated that substrate scope tables from journal articles, known for being small and biased towards high-yielding examples, are in fact valuable sources of information for molecular representation learning. This is the first approach to utilize these groupings

as an information source and to demonstrate their correlation with reactivity features and expert intuition. By learning patterns of human bias, ContraScope extracts the expert knowledge and reactivity trends underlying this bias. We illustrate this both qualitatively and quantitatively, showing how our learned embeddings align with established physical organic chemistry descriptors.

While the applications showcased in this paper—such as the supervised learning of reaction yields and regioselectivity, along with the diverse substrate selection for experimental design—underscore the versatility of our learned embeddings, it is important to recognize that they serve as proofs-of-concept; our embedding approach, like any other, is not a universal solution. For instance, a nearest neighbor regression analysis of yield across all training substrate scopes indicates that two molecules with similar embeddings do not necessarily exhibit similar yields, which is similar to other common features (Figure S14). Moreover, substrates may achieve low reaction yields as a result of side reactions rather than inherent lack of reactivity at the C–X motif, which would not be captured in our approach; this may partially explain the observed lack of correlation for structurally complex aryl halides in the chemistry informer library⁶¹ (Figure S15). Additionally, our operational presumption that all unobserved molecules in scope tables are less similar is a simplification that holds true only on average. From a practical standpoint, our loss function combines a term for anchor-positive loss with one for anchor-negative loss in a linear manner, which poses challenges in training both loss terms simultaneously and can at times lead to instability in the training process. All those challenges underscore that understanding and predicting chemical reactivity remains an open question, highlighting the need for further refinement of pre-training strategies and loss functions, specially designed for chemical data.

Overall, our work presents a conceptual framework for pre-training reactivity-aligned atomic representations from the selective bias inherent in substrate scope tables. With the availability of our code and pre-trained model, this methodology can be applied to new reaction systems and to develop novel chemical embeddings for diverse molecular classes. We are

optimistic that our framework will serve as a catalyst for further advancements in the field, fostering more targeted and functionality-aligned approaches in molecular representation learning.

Methods

Substrate data for training In this study, we focus on aryl halides. We partitioned reactions recorded in the CAS Content Collection between 2010 and 2015 into substrate scope tables comprised of reactions from the same publication source, wherein the only variable is a single aryl halide substrate. All other substrates and recorded conditions remained constant within each scope. Importantly, all reactions take place at an aryl C–X bond. Scopes with fewer than 5 reactions and reactions without recorded yields were excluded. This led to our final training dataset of 20,798 distinct aryl halides, covering 64,192 reactions and 6,919 substrate scopes.

Calculation of reactivity descriptors All reactivity descriptors are calculated by density-functional theory (DFT) with Gaussian 16⁵⁶ via an automated descriptor generation pipeline built on top of AutoQChem.⁵⁷ For each unique aryl halide, up to 20 conformers were generated using RDKit’s ETKDG algorithm⁶³ and optimized using the MMFF94 force field.⁶⁴ To reduce computational overhead when using DFT, the lowest-energy conformer for each molecule was then selected via GFN2-xTB.⁶⁵ Any conformer for which any energy calculation did not converge was discarded. The lowest-energy conformer for each aryl halide then underwent geometry optimization and frequency calculations in Gaussian 16 using the B3LYP functional^{66,67} with the 6-31G*⁶⁸ basis set. Atoms with atomic number > 35 use the LANL2DZ⁶⁹ basis set instead. This led to the generation of 25 molecule-level descriptors per molecule (energies, energy corrections, dipole moment, HOMO/LUMO energies, electronegativity, etc.) and 19 atom-level descriptors per atom per molecule (buried volume, partial charges, NMR shielding constants, etc.), and the relevant atom-level descriptors for

the reactive C–X bond were extracted. The full set of descriptors is listed in Figures S5 and S6.

Featurization and network architecture We adopt a graph representation of molecules and the graph isomorphism network (GIN)⁴⁸ for modeling them. GIN is one kind of graph neural network that updates the representation of each atom over multiple iterations as follows:

$$h_v^{(k)} = \text{MLP}^{(k)} \left((1 + \epsilon^{(k)}) h_v^{(k-1)} + \sum_{u \in \mathcal{N}(v)} h_u^{(k-1)} \right) \quad (2)$$

where $h_v^{(k)}$ is the atom representation of atom v at k -th layer, $\mathcal{N}(v)$ is a set of atoms that are covalently bonded to atom v . ϵ is a learnable scalar and MLP represents a multi-layer perceptrons model. Utilizing this atom-wise message-passing schema, GIN is able to propagate and aggregate information through network layers to embed atoms in a manner that respects graph isomorphism. After hyper-parameter tuning, the network has a hidden dimension of 64, and message passing layer of 5. For the initial representation of molecular graphs, we use minimal featurizations⁴⁹ that include atomic numbers, total degrees, formal charges, chiral tags, number of hydrogens, hybridization types, aromaticity, mass, and bond types.

Substrate scope contrastive learning The training of substrate scope contrastive learning involves triplet sampling, computing loss function depicted in Eq. 1, and back-propagation. Within an epoch, each aryl halide serves as the anchor molecule once and 16 triplets are sampled randomly for each anchor. We exclude all cases with identical molecules sampled in one triplet. Based on empirical performance, we adopt a distance metric called signal-noise ratio distance (SNR),⁷⁰ defined as follows:

$$d(m_i, m_j) = d_{\text{SNR}}(f_i, f_j) = \frac{\text{var}(f_i)}{\text{var}(f_j - f_i)} \quad (3)$$

where f_i, f_j is the embedding of aryl halide m_i, m_j . We take the representation of an aryl halide's reactivity, f_i , to be the atom-level feature of the carbon atom in the reactive C–

X bond from the GIN. The Adam⁷¹ optimization algorithm is used for stochastic gradient descent.

To monitor the model performance during the training, we collected four datasets evaluating the ability of the learned embeddings to adapt to downstream tasks: 15 aryl bromides with associated reaction yields for a cross-coupling reaction,³ 61 mono-substituted aryl halides with Hammett constants,⁷² and calculated Mulliken charges and NMR shielding constants. At the end of each epoch, we used the learned embeddings as features, selected the most informative features indicated by mutual information⁷³ and the most correlated features indicated by the Pearson correlation coefficient, and trained a k-Nearest Neighbor (k-NN) to predict the reaction yields, Hammett constants, Mulliken charges and NMR shielding constants. We evaluated the r^2 ⁷⁴ of prediction with leave-one-out cross-validation on the four validation datasets and summed them up as a single scalar value for monitoring the training process and determining the time of early termination of the training. Training curves are shown in Figure S4.

Validation with traditional reactivity descriptors To validate if the embedding learns anything about reactivity, we selected the most commonly used 500 aryl halides from the training set and supplemented them with p-, m-, and o-substituted aryl halides with common functional groups selected from a study of Hammett constants,⁵⁵ providing a dataset with 762 aryl halides in total. Reactivity descriptors were calculated for this set of 762 molecules. For each molecule, we computed RDKit2D descriptors⁷⁵ and Morgan fingerprints⁵⁹ (1024 bits, radius 2) using Therapeutic Data Commons (TDC).⁷⁶ As other pre-trained representation baselines, we employed a pre-trained GIN with node-masking²⁶ and ChemGPT³⁷ with 4.7 M parameters, available from Deep Graph Library (DGL)⁷⁷ and HuggingFace,⁷⁸ interfaced via molfeat.⁷⁹ For Figure 3B, we assessed k-Nearest Neighbor (k-NN), linear regression, L1 and L2 norm, and support vector machine (SVM)⁵³ on each embedding, reporting the best r^2 performance. Performance validation used a 3-fold cross-validation, with mean and range of r^2 for the top model is shown in the bar plot. In Figure 3C, we trained an SVM model⁵³ on

a specific subset of the training data, using the remaining data for performance evaluation. This was repeated thrice with independent seeds, reporting the mean and range. For all machine learning models, we adopt the implementation from scikit-learn.⁸⁰

Downstream applications For yield prediction, we used 15 aryl bromides with experimental yields,³ employing leave-one-out validation. Molecular embeddings were calculated by the ContraScope-trained GIN as previously described, and other common featurization was calculated using the same pipeline as above. For all embeddings, we selected features that are the most informative or most correlated with the prediction targets indicated by mutual information⁷³ and Pearson correlation coefficient, and tested models including k-NN, linear regression, L1 and L2 norms, SVM, random forest, and MLP, reported the highest validation r^2 . The r^2 value of the univariate model using electronegativity is from the original study³ for comparison. For regioselectivity prediction, training data comprised penta- and trifluoronitrobenzenes to predict reactivity at C-F sites in tetrafluoronitrobenzenes.⁶⁰ Each C-F site was encoded and labeled with its yield if reactive, or zero otherwise. During inference, all C-F sites were encoded, and the model predicted each site's reactivity as a binary classification task. For substrate scope design, we used the same purchasable aryl bromides as Kariofillis et al. and encoded them for K-Means clustering into 15 groups, selecting a representative from each. All machine learning models were implemented using scikit-learn.⁸⁰

Acknowledgement

This research was supported by the National Science Foundation under Grant No. CHE-2144153. W.G. is supported by the Google PhD Fellowship. W.G. and P.R. received additional support from the MIT-Takeda Fellowship program. R.S. received additional support from the MIT UROP Office. We thank CAS for providing data from the CAS Content Collection needed to enable this study and Yitong Tseo for preliminary work exploring the

dataset.

References

- (1) Hickey, D. P.; Schiedler, D. A.; Matanovic, I.; Doan, P. V.; Atanassov, P.; Minteer, S. D.; Sigman, M. S. Predicting electrocatalytic properties: modeling structure–activity relationships of nitroxyl radicals. *Journal of the American Chemical Society* **2015**, *137*, 16179–16186.
- (2) Santiago, C. B.; Guo, J.-Y.; Sigman, M. S. Predictive and mechanistic multivariate linear regression models for reaction development. *Chemical science* **2018**, *9*, 2398–2412.
- (3) Kariofillis, S. K.; Jiang, S.; Zuranski, A. M.; Gandhi, S. S.; Martinez Alvarado, J. I.; Doyle, A. G. Using data science to guide aryl bromide substrate scope analysis in a Ni/photoredox-catalyzed cross-coupling with acetals as alcohol-derived radical sources. *Journal of the American Chemical Society* **2022**, *144*, 1045–1055.
- (4) Tang, T.; Hazra, A.; Min, D. S.; Williams, W. L.; Jones, E.; Doyle, A. G.; Sigman, M. S. Interrogating the Mechanistic Features of Ni (I)-Mediated Aryl Iodide Oxidative Addition Using Electroanalytical and Statistical Modeling Techniques. *Journal of the American Chemical Society* **2023**, *145*, 8689–8699.
- (5) Jordan, M. I.; Mitchell, T. M. Machine learning: Trends, perspectives, and prospects. *Science* **2015**, *349*, 255–260.
- (6) LeCun, Y.; Bengio, Y.; Hinton, G. Deep learning. *nature* **2015**, *521*, 436–444.
- (7) Schwaller, P.; Vaucher, A. C.; Laino, T.; Reymond, J.-L. Prediction of chemical reaction yields using deep learning. *Machine learning: science and technology* **2021**, *2*, 015016.

- (8) Gallegos, L. C.; Luchini, G.; St. John, P. C.; Kim, S.; Paton, R. S. Importance of engineered and learned molecular representations in predicting organic reactivity, selectivity, and chemical properties. *Accounts of Chemical Research* **2021**, *54*, 827–836.
- (9) Probst, D.; Schwaller, P.; Reymond, J.-L. Reaction classification and yield prediction using the differential reaction fingerprint DRFP. *Digital discovery* **2022**, *1*, 91–97.
- (10) Huh, M.; Agrawal, P.; Efros, A. A. What makes ImageNet good for transfer learning? *arXiv preprint arXiv:1608.08614* **2016**,
- (11) Devlin, J.; Chang, M.-W.; Lee, K.; Toutanova, K. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805* **2018**,
- (12) Radford, A.; Narasimhan, K.; Salimans, T.; Sutskever, I., et al. Improving language understanding by generative pre-training. **2018**,
- (13) Bommasani, R.; Hudson, D. A.; Adeli, E.; Altman, R.; Arora, S.; von Arx, S.; Bernstein, M. S.; Bohg, J.; Bosselut, A.; Brunskill, E., et al. On the opportunities and risks of foundation models. *arXiv preprint arXiv:2108.07258* **2021**,
- (14) He, K.; Chen, X.; Xie, S.; Li, Y.; Dollár, P.; Girshick, R. Masked autoencoders are scalable vision learners. Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2022; pp 16000–16009.
- (15) Chen, W.-Y.; Liu, Y.-C.; Kira, Z.; Wang, Y.-C. F.; Huang, J.-B. A closer look at few-shot classification. *arXiv preprint arXiv:1904.04232* **2019**,
- (16) Wang, Y.; Yao, Q.; Kwok, J. T.; Ni, L. M. Generalizing from a few examples: A survey on few-shot learning. *ACM computing surveys (csur)* **2020**, *53*, 1–34.
- (17) Koch, G.; Zemel, R.; Salakhutdinov, R., et al. Siamese neural networks for one-shot image recognition. ICML deep learning workshop. 2015.

- (18) Snell, J.; Swersky, K.; Zemel, R. Prototypical networks for few-shot learning. *Advances in neural information processing systems* **2017**, *30*.
- (19) Sung, F.; Yang, Y.; Zhang, L.; Xiang, T.; Torr, P. H.; Hospedales, T. M. Learning to compare: Relation network for few-shot learning. Proceedings of the IEEE conference on computer vision and pattern recognition. 2018; pp 1199–1208.
- (20) Li, H.; Eigen, D.; Dodge, S.; Zeiler, M.; Wang, X. Finding task-relevant features for few-shot learning by category traversal. Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2019; pp 1–10.
- (21) Brown, T.; Mann, B.; Ryder, N.; Subbiah, M.; Kaplan, J. D.; Dhariwal, P.; Neelakantan, A.; Shyam, P.; Sastry, G.; Askell, A., et al. Language models are few-shot learners. *Advances in neural information processing systems* **2020**, *33*, 1877–1901.
- (22) Ouyang, L.; Wu, J.; Jiang, X.; Almeida, D.; Wainwright, C.; Mishkin, P.; Zhang, C.; Agarwal, S.; Slama, K.; Ray, A., et al. Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems* **2022**, *35*, 27730–27744.
- (23) Liu, P.; Yuan, W.; Fu, J.; Jiang, Z.; Hayashi, H.; Neubig, G. Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing. *ACM Computing Surveys* **2023**, *55*, 1–35.
- (24) Rao, R.; Meier, J.; Sercu, T.; Ovchinnikov, S.; Rives, A. Transformer protein language models are unsupervised structure learners. International Conference on Learning Representations. 2020.
- (25) Evans, R.; O'Neill, M.; Pritzel, A.; Antropova, N.; Senior, A.; Green, T.; Žídek, A.; Bates, R.; Blackwell, S.; Yim, J., et al. Protein complex prediction with AlphaFold-Multimer. *biorxiv* **2021**, 2021–10.

- (26) Hu, W.; Liu, B.; Gomes, J.; Zitnik, M.; Liang, P.; Pande, V.; Leskovec, J. Strategies For Pre-training Graph Neural Networks. International Conference on Learning Representations (ICLR). 2020.
- (27) Wang, H.; Li, W.; Jin, X.; Cho, K.; Ji, H.; Han, J.; Burke, M. D. Chemical-reaction-aware molecule representation learning. *arXiv preprint arXiv:2109.09888* **2021**,
- (28) Liu, S.; Wang, H.; Liu, W.; Lasenby, J.; Guo, H.; Tang, J. Pre-training molecular graph representation with 3d geometry. *arXiv preprint arXiv:2110.07728* **2021**,
- (29) Wang, Y.; Wang, J.; Cao, Z.; Barati Farimani, A. Molecular contrastive learning of representations via graph neural networks. *Nature Machine Intelligence* **2022**, *4*, 279–287.
- (30) Stärk, H.; Beaini, D.; Corso, G.; Tossou, P.; Dallago, C.; Günemann, S.; Liò, P. 3d infomax improves gnns for molecular property prediction. International Conference on Machine Learning. 2022; pp 20479–20502.
- (31) Sun, R.; Dai, H.; Yu, A. W. Does GNN Pretraining Help Molecular Representation? *Advances in Neural Information Processing Systems* **2022**, *35*, 12096–12109.
- (32) Rong, Y.; Bian, Y.; Xu, T.; Xie, W.; Wei, Y.; Huang, W.; Huang, J. Self-supervised graph transformer on large-scale molecular data. *Advances in Neural Information Processing Systems* **2020**, *33*, 12559–12571.
- (33) Weininger, D. SMILES, a chemical language and information system. 1. Introduction to methodology and encoding rules. *Journal of chemical information and computer sciences* **1988**, *28*, 31–36.
- (34) Honda, S.; Shi, S.; Ueda, H. R. Smiles transformer: Pre-trained molecular fingerprint for low data drug discovery. *arXiv preprint arXiv:1911.04738* **2019**,

- (35) Schwaller, P.; Laino, T.; Gaudin, T.; Bolgar, P.; Hunter, C. A.; Bekas, C.; Lee, A. A. Molecular transformer: a model for uncertainty-calibrated chemical reaction prediction. *ACS central science* **2019**, *5*, 1572–1583.
- (36) Chithrananda, S.; Grand, G.; Ramsundar, B. ChemBERTa: large-scale self-supervised pretraining for molecular property prediction. *arXiv preprint arXiv:2010.09885* **2020**,
- (37) Frey, N. C.; Soklaski, R.; Axelrod, S.; Samsi, S.; Gomez-Bombarelli, R.; Coley, C. W.; Gadepally, V. Neural scaling of deep chemical models. *Nature Machine Intelligence* **2023**, 1–9.
- (38) Wang, X.; Zhao, H.; Tu, W.-w.; Yao, Q. Automated 3D pre-training for molecular property prediction. Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining. 2023; pp 2419–2430.
- (39) Tu, Z.; Stuyver, T.; Coley, C. W. Predictive chemistry: machine learning for reaction deployment, reaction development, and reaction discovery. *Chemical Science* **2023**,
- (40) Schultz, M.; Joachims, T. Learning a distance metric from relative comparisons. *Advances in neural information processing systems* **2003**, *16*.
- (41) Hoffer, E.; Ailon, N. Deep metric learning using triplet network. Similarity-Based Pattern Recognition: Third International Workshop, SIMBAD 2015, Copenhagen, Denmark, October 12–14, 2015. Proceedings 3. 2015; pp 84–92.
- (42) Khosla, P.; Teterwak, P.; Wang, C.; Sarna, A.; Tian, Y.; Isola, P.; Maschinot, A.; Liu, C.; Krishnan, D. Supervised contrastive learning. *Advances in neural information processing systems* **2020**, *33*, 18661–18673.
- (43) Kozlowski, M. C. On the Topic of Substrate Scope. 2022.

- (44) Guastavino, J. F.; Buden, M. E.; Rossi, R. A. Room-Temperature and Transition-Metal-Free Mizoroki–Heck-type Reaction. Synthesis of E-Stilbenes by Photoinduced C–H Functionalization. *The Journal of Organic Chemistry* **2014**, *79*, 9104–9111.
- (45) Movahed, S. K.; Dabiri, M.; Bazgir, A. A one-step method for preparation of Cu@ Cu₂O nanoparticles on reduced graphene oxide and their catalytic activities in N-arylation of N-heterocycles. *Applied Catalysis A: General* **2014**, *481*, 79–88.
- (46) Roughley, S. D.; Jordan, A. M. The medicinal chemist's toolbox: an analysis of reactions used in the pursuit of drug candidates. *Journal of medicinal chemistry* **2011**, *54*, 3451–3479.
- (47) Brown, D. G.; Bostrom, J. Analysis of past and present synthetic methodologies on medicinal chemistry: where have all the new reactions gone? Miniperspective. *Journal of medicinal chemistry* **2016**, *59*, 4443–4458.
- (48) Xu, K.; Hu, W.; Leskovec, J.; Jegelka, S. How powerful are graph neural networks? *arXiv preprint arXiv:1810.00826* **2018**,
- (49) Yang, K.; Swanson, K.; Jin, W.; Coley, C.; Eiden, P.; Gao, H.; Guzman-Perez, A.; Hopper, T.; Kelley, B.; Mathea, M., et al. Analyzing learned molecular representations for property prediction. *Journal of chemical information and modeling* **2019**, *59*, 3370–3388.
- (50) Hinton, G. E.; Roweis, S. Stochastic neighbor embedding. *Advances in neural information processing systems* **2002**, *15*.
- (51) Chen, X.; Reynolds, C. H. Performance of similarity measures in 2D fragment-based similarity searching: comparison of structural descriptors and similarity coefficients. *Journal of chemical information and computer sciences* **2002**, *42*, 1407–1414.

- (52) Bajusz, D.; Rácz, A.; Héberger, K. Why is Tanimoto index an appropriate choice for fingerprint-based similarity calculations? *Journal of cheminformatics* **2015**, *7*, 1–13.
- (53) Hearst, M. A.; Dumais, S. T.; Osuna, E.; Platt, J.; Scholkopf, B. Support vector machines. *IEEE Intelligent Systems and their applications* **1998**, *13*, 18–28.
- (54) Kohn, W.; Sham, L. Density functional theory. CONFERENCE PROCEEDINGS-ITALIAN PHYSICAL SOCIETY. 1996; pp 561–572.
- (55) Hansch, C.; Leo, A.; Taft, R. A survey of Hammett substituent constants and resonance and field parameters. *Chemical reviews* **1991**, *91*, 165–195.
- (56) Frisch, M. J. et al. Gaussian[®]16 Revision C.01. 2016; Gaussian Inc. Wallingford CT.
- (57) Żurański, A. M.; Wang, J. Y.; Shields, B. J.; Doyle, A. G. Auto-QChem: an automated workflow for the generation and storage of DFT calculations for organic molecules. *Reaction Chemistry & Engineering* **2022**, *7*, 1276–1284.
- (58) Ashton, M.; Barnard, J.; Casset, F.; Charlton, M.; Downs, G.; Gorse, D.; Holliday, J.; Lahana, R.; Willett, P. Identification of diverse database subsets using property-based and fragment-based molecular descriptions. *Quantitative Structure-Activity Relationships* **2002**, *21*, 598–604.
- (59) Rogers, D.; Hahn, M. Extended-connectivity fingerprints. *Journal of chemical information and modeling* **2010**, *50*, 742–754.
- (60) Cargill, M. R.; Sandford, G.; Tadeusiak, A. J.; Yufit, D. S.; Howard, J. A.; Kilickiran, P.; Nelles, G. Palladium-catalyzed C- F activation of polyfluoronitrobenzene derivatives in Suzuki- Miyaura coupling reactions. *The Journal of Organic Chemistry* **2010**, *75*, 5860–5866.
- (61) Kutchukian, P. S.; Dropinski, J. F.; Dykstra, K. D.; Li, B.; DiRocco, D. A.; Streckfuss, E. C.; Campeau, L.-C.; Cernak, T.; Vachal, P.; Davies, I. W., et al. Chemistry

- informer libraries: a chemoinformatics enabled approach to evaluate and advance synthetic methods. *Chemical science* **2016**, *7*, 2604–2613.
- (62) Lloyd, S. Least squares quantization in PCM. *IEEE transactions on information theory* **1982**, *28*, 129–137.
- (63) Riniker, S.; Landrum, G. A. Better informed distance geometry: using what we know to improve conformation generation. *Journal of chemical information and modeling* **2015**, *55*, 2562–2574.
- (64) Halgren, T. A. Merck molecular force field. I. Basis, form, scope, parameterization, and performance of MMFF94. *Journal of computational chemistry* **1996**, *17*, 490–519.
- (65) Bannwarth, C.; Ehlert, S.; Grimme, S. GFN2-xTB—An accurate and broadly parametrized self-consistent tight-binding quantum chemical method with multipole electrostatics and density-dependent dispersion contributions. *Journal of chemical theory and computation* **2019**, *15*, 1652–1671.
- (66) Becke, A. D. Density-functional thermochemistry. I. The effect of the exchange-only gradient correction. *The Journal of chemical physics* **1992**, *96*, 2155–2160.
- (67) Lee, C.; Yang, W.; Parr, R. G. Development of the Colle-Salvetti correlation-energy formula into a functional of the electron density. *Physical review B* **1988**, *37*, 785.
- (68) Petersson, a.; Bennett, A.; Tensfeldt, T. G.; Al-Laham, M. A.; Shirley, W. A.; Mantzaris, J. A complete basis set model chemistry. I. The total energies of closed-shell atoms and hydrides of the first-row elements. *The Journal of chemical physics* **1988**, *89*, 2193–2218.
- (69) Hay, P. J.; Wadt, W. R. Ab initio effective core potentials for molecular calculations. Potentials for the transition metal atoms Sc to Hg. *The Journal of chemical physics* **1985**, *82*, 270–283.

- (70) Yuan, T.; Deng, W.; Tang, J.; Tang, Y.; Chen, B. Signal-to-noise ratio: A robust distance metric for deep metric learning. Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2019; pp 4815–4824.
- (71) Kingma, D. P.; Ba, J. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980* **2014**,
- (72) Hammett, L. P. The effect of structure upon the reactions of organic compounds. Benzene derivatives. *Journal of the American Chemical Society* **1937**, *59*, 96–103.
- (73) Shannon, C. E. A mathematical theory of communication. *The Bell system technical journal* **1948**, *27*, 379–423.
- (74) Wright, S. Correlation and causation. **1921**,
- (75) Landrum, G., et al. RDKit: A software suite for cheminformatics, computational chemistry, and predictive modeling. *Greg Landrum* **2013**, *8*, 31.
- (76) Huang, K.; Fu, T.; Gao, W.; Zhao, Y.; Roohani, Y. H.; Leskovec, J.; Coley, C. W.; Xiao, C.; Sun, J.; Zitnik, M. Therapeutics Data Commons: Machine Learning Datasets and Tasks for Drug Discovery and Development. Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 1). 2021.
- (77) Wang, M.; Zheng, D.; Ye, Z.; Gan, Q.; Li, M.; Song, X.; Zhou, J.; Ma, C.; Yu, L.; Gai, Y., et al. Deep graph library: A graph-centric, highly-performant package for graph neural networks. *arXiv preprint arXiv:1909.01315* **2019**,
- (78) Wolf, T.; Debut, L.; Sanh, V.; Chaumond, J.; Delangue, C.; Moi, A.; Cistac, P.; Rault, T.; Louf, R.; Funtowicz, M., et al. Huggingface’s transformers: State-of-the-art natural language processing. *arXiv preprint arXiv:1910.03771* **2019**,
- (79) Noutahi, E.; Wognum, C.; Mary, H.; Hounwanou, H.; Kovary, K. M.; Gilmour, D.;

- Burns, J.; St-Laurent, J.; D.; Maheshkar, S.; rbyrne momatx, datamol-io/molfeat: 0.9.4 (0.9.4). 2023; Zenodo. <https://doi.org/10.5281/zenodo.8373019>.
- (80) Pedregosa, F. et al. Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research* **2011**, *12*, 2825–2830.

Supporting Information

Substrate Scope Contrastive Learning: Repurposing Artificial Bias to Learn Atomic Representations

Wenhao Gao, Priyanka Raghavan, Ron Shprints and Connor W. Coley*

Department of Chemical Engineering, MIT, Cambridge, MA 02139

E-mail: ccoley@mit.edu

Code and Data Availability

All code necessary to reproduce this work can be found at https://github.com/wenhao-gao/substrate_scope_contrastive_learning/tree/main.

The misalignment of current pre-training methods and functionalities

Recent years have witnessed significant advancements in representation learning for molecules, accompanied by a diverse array of pre-training strategies. These strategies have emphasized various aspects, including chemical valence,²⁶ structural similarity,²⁹ and conformational information.^{28,30} Despite these developments, a consistent or marked enhancement in performance for downstream tasks remains elusive, as noted in Sun et al. (2022).³¹ A contributing factor to this challenge may be the misalignment between the objectives of these learning models and the specific requirements of their target applications, particularly in modeling molecular functionality.

Focusing on graph neural networks, early methodologies primarily targeted node or contextual prediction, as detailed in Hu et al. (2020).²⁶ These approaches often classified atoms

or groups with identical chemical valences as analogous. This technique, however, shows limitations, as evidenced in the molecules depicted in Figure S1. Here, the pre-training method erroneously identifies distinct functional groups as identical, overlooking substantial differences in their chemical functionalities. Similarly, traditional contrastive learning, which views structurally similar molecules as comparable in the embedding space,²⁹ falls short. The assumption that consistent valence bonds or structural resemblance equates to analogous molecular properties is often flawed, as significant variances may arise. While the integration of 3D conformational data offers benefits for properties reliant on specific conformations, such as those derived from quantum chemical computations,^{28,30} its utility diminishes in broader biological or reaction contexts. Additionally, language model pre-training that focuses on string representations like SMILES³³ tends to prioritize syntax comprehension over understanding the intrinsic molecular significance.

This view of the landscape of pre-training approaches for small molecules, intended to be used for the prediction of chemical reactivity but only using information about structure and conformation, served as inspiration for this study. ContraScope is a pre-training approach that is fundamentally based on chemical reactivity.

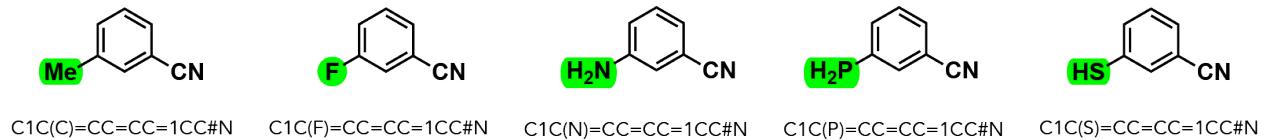
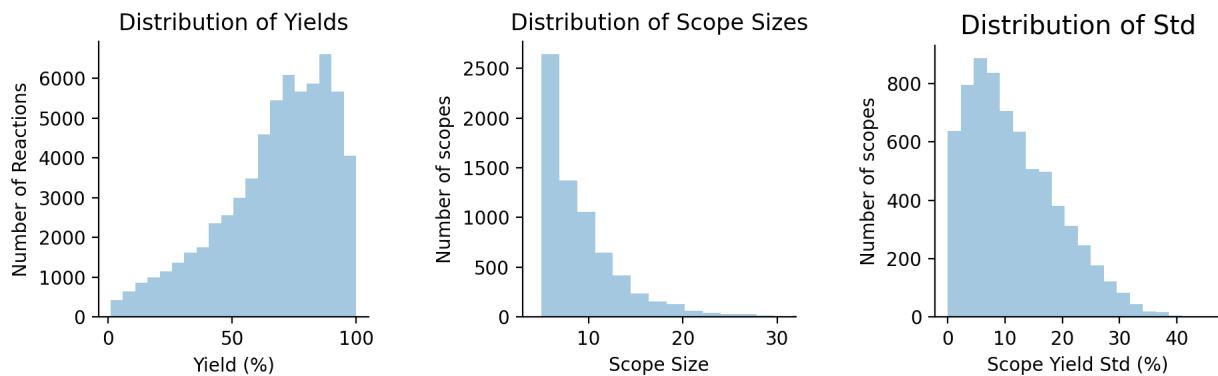


Figure S1: Illustration of various substituted benzonitriles. Despite the structural similarity of the benzonitrile backbone across all molecules, the diverse electronic properties of the substituents significantly influence the molecular functionality. This exemplifies the limitation of some pre-training methods in graph neural networks that may not distinguish between these functional nuances. The corresponding SMILES notations are provided below each molecule, indicating that the highlighted substituents are also interchangeable in string representation, which also brings into question the reasonableness of string-based pre-training methodologies.

Statistics of the dataset

In Figure S2, we present statistics of the substrate scope dataset we derive from the CAS Content CollectionTM, which includes the size distribution of the substrate scopes, the yield values' distribution for all reactions utilized in the training set, and the distribution of the yield standard deviation within these scopes. The data reveal a predominant trend of small substrate scopes, with the majority comprising fewer than 20 substrates. Additionally, there is a discernible skew towards higher yields within the dataset indicating the artificial selective bias that hindered typical supervised machine learning.



- (a) The distribution of the yields of reactions used in training.
(b) The distribution of the sizes of substrate scopes.
(c) The distribution of the yields' standard deviation within substrate scopes.

Figure S2: The statistics of the substrate scope dataset used for training.

In Figure S3, we show the cumulative coverage of reactions by the top-k most frequently occurring aryl halides. Notably, the data indicate that the 500 aryl halides with the highest frequency of occurrence account for more than half of the total reactions in the training set. This observation underscores the existence of a frequently utilized subset of aryl halides and substantiates our methodological choice to focus on a select group of these compounds for in-depth analysis.

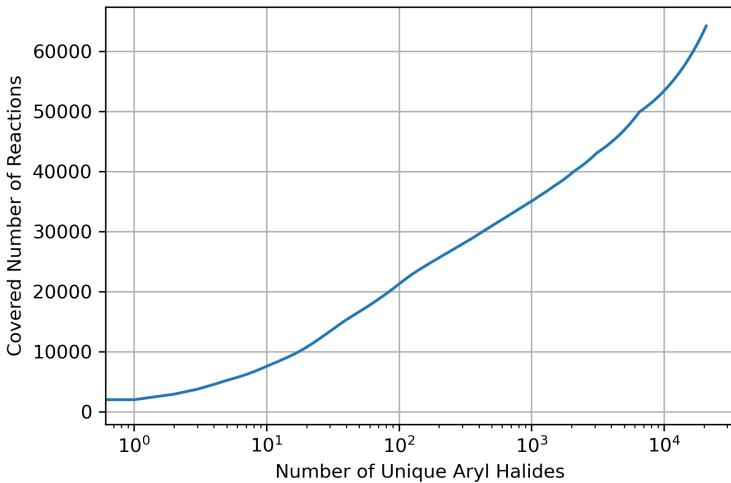


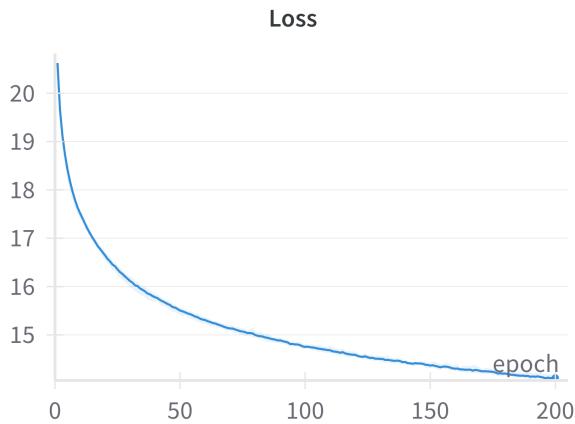
Figure S3: Illustration of the coverage of reactions by the top-k most frequently occurring aryl halides. The x-axis, which is log-scaled for better visual comprehension, represents the k aryl halides ranked by frequency of appearance. It is observed that the 500 most prevalent aryl halides account for over half of the reactions in the dataset.

Learning curves and hyper-parameter tuning

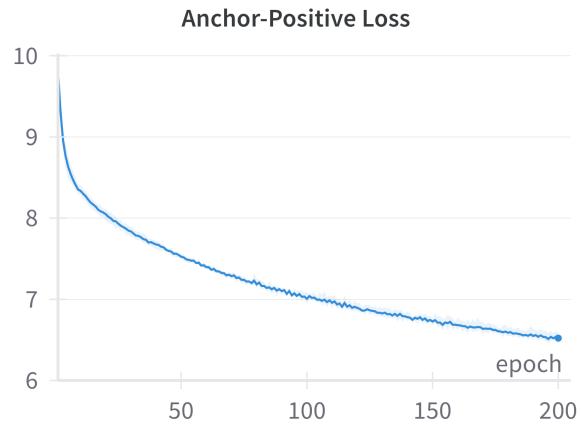
The GIN network was trained with our substrate scope contrastive loss, as described in Equation 1. The value of the total loss, anchor-positive term, and anchor-negative term are depicted in Figure S4(a-c). The cumulative gradient of the total loss across all trainable parameters is illustrated in Figure S4(d). As described in the Method section, under the *Substrate scope contrastive learning* subsection, we constructed four validation tasks to monitor the training process. This monitoring involved evaluating the network’s ability to predict reaction yields, Hammett constants, Mulliken charges, and NMR shielding constants. The summation of the coefficient of determination (r^2) values across these four validation tasks, as well as the Pearson correlation coefficient between the predicted targets and their most correlated features, are presented in Figure S4(e-f). Overall, we can see the r^2 and the Pearson correlation coefficient plateau after 50 epochs and oscillate after that. We terminated the training at the 56th epoch, which reached the highest aggregate r^2 value.

We tuned the hyper-parameter of the network architecture and the training details to

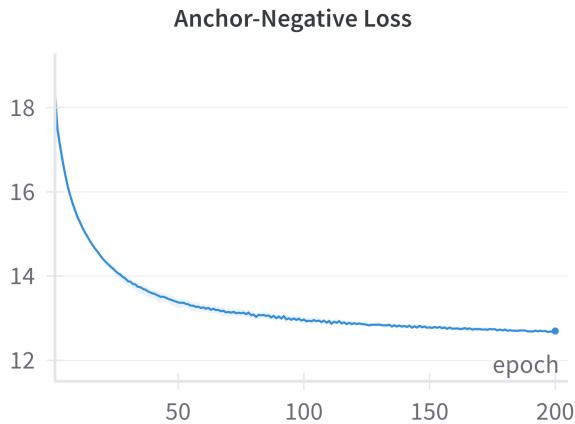
maximize the summation of r^2 across these four validation tasks. The resulting GIN network has 64 hidden channels, 5 layers. An Adam optimizer was used with a momentum of 0.9, an initial learning rate of 0.00005, and a learning rate decay of 0.999. 16 triplets are sampled for each anchor molecule and a batch size of 4096 was used for training. In Equation 1, we used γ of 4.022879258650723, β of 0.6120957227224214, and M of 1.906408074987083.



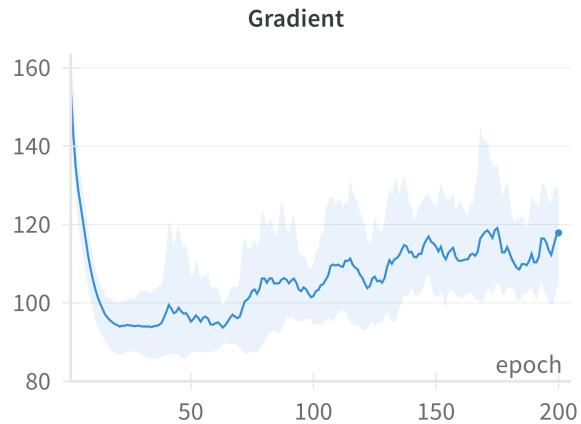
(a) Total substrate scope contrastive loss.



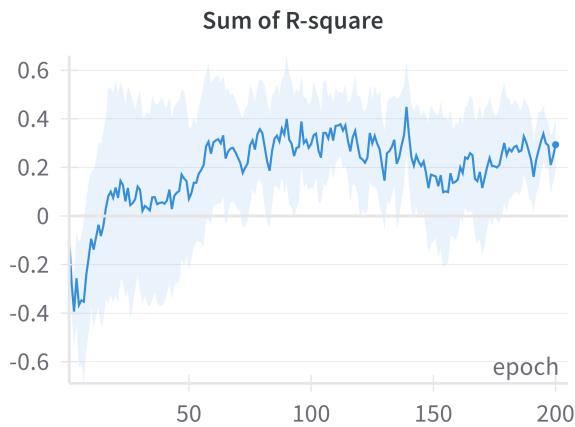
(b) The value of the anchor-positive term.



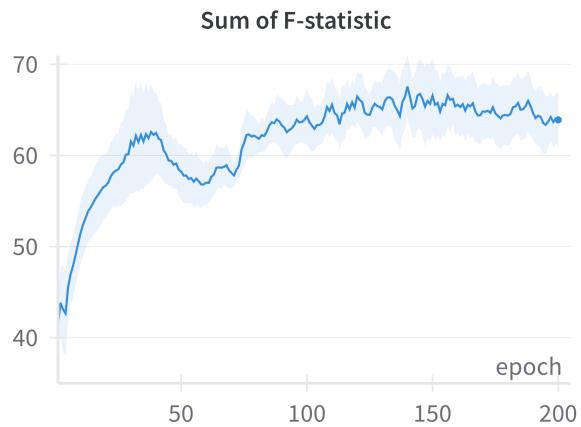
(c) The value of the anchor-negative term.



(d) The sum of the gradient of total loss.



(e) The summation of r^2 across the four validation tasks.



(f) The summation of the Pearson correlation coefficient across the four validation tasks.

Figure S4: The learning curves of the pre-training process. The mean and range of four curves with distinct random seeds are reported.

Details of reactivity descriptors

In the following tables, we present descriptions of all reactivity descriptors extracted from DFT-level calculations carried out on the aryl halides.

Global (Molecule-Level) Descriptors	
Descriptor	Explanation
number_of_atoms	Number of atoms in the molecule
charge	Overall charge on the molecule
multiplicity	Spin multiplicity of the molecule
dipole	Dipole moment (net polarity) of the molecule
molar_mass	Molecular weight of the molecule
electronic_spatial_extent	Sum of Cartesian displacements for the electronic contributions to the wavefunction. Measures the extent of electronic density.
E_scf	Total electronic energy of the molecule, corresponding to the molecule's electrons and nuclei at infinite separation
zero_point_correction	Zero-point vibrational energy arising from all vibrational modes of the molecule. Note that the correction value is given prior to conversion to a per mol basis
E_zpe	Sum of electronic and zero-point energies
{E/H/G}_thermal_correction	Thermal corrections to energies/enthalpies/free energies, arising from electronic, vibrational, translational, and rotational contributions, and given at 298.15 K
E/H/G	Sum of electronic and thermal energies/enthalpies/free energies
homo_energy	Energy of the HOMO (highest-occupied molecular orbital)
lumo_energy	Energy of the LUMO (lowest-unoccupied molecular orbital)
electronegativity	Electronegativity of the molecule
hardness	Hardness of the molecule

Figure S5: All computed global (molecule-level) descriptors using autoqchem,⁵⁷ and an explanation of each.

Local (Atom-Level) Descriptors	
Descriptor	Explanation
VBur	Occupied volume fraction of the atom within a sphere sized to its Van der Waals radius, given the atom's position. Note that this is a calculated value given the optimized geometry coordinates
Mulliken_charge	Atomic charge obtained from Mulliken population analysis (derived from original basis functions)
APT_charge	Atomic charge obtained from generalized atomic polar tensors. Note that these are obtained from running frequency calculations
NPA_charge	Atomic charges obtained from natural population analysis (derived from natural atomic orbitals (NAOs)). Note that this represents the summed NAO populations on the atom, subtracted from the nuclear charge
NPA_{core/valence/Rydberg}	Populations of core/valence/Rydberg-shell NAOs for each atom
NPA_total	Sum of core, valence, and Rydberg NAO populations on each atom
NMR_shift	Isotropic nuclear magnetic shielding tensor computed by the gauge-independent atomic orbital (GIAO) method. Note that this is reported in ppm
NMR_anisotropy	Anisotropic nuclear magnetic shielding tensor computed by the GIAO method. Note that this is reported in ppm

Figure S6: All computed local (atom-level) descriptors using autoqchem,⁵⁷ and an explanation of each.

Additional Results

Intuitive investigation of the learned embeddings

To offer an intuitive examination of the bit-level details of the embeddings, we illustrate the bit value of the corresponding position in the embeddings with a heatmap and their hierarchical clustering in Figure S7. We analyzed the same set of aryl bromides as Figure 2B-C.

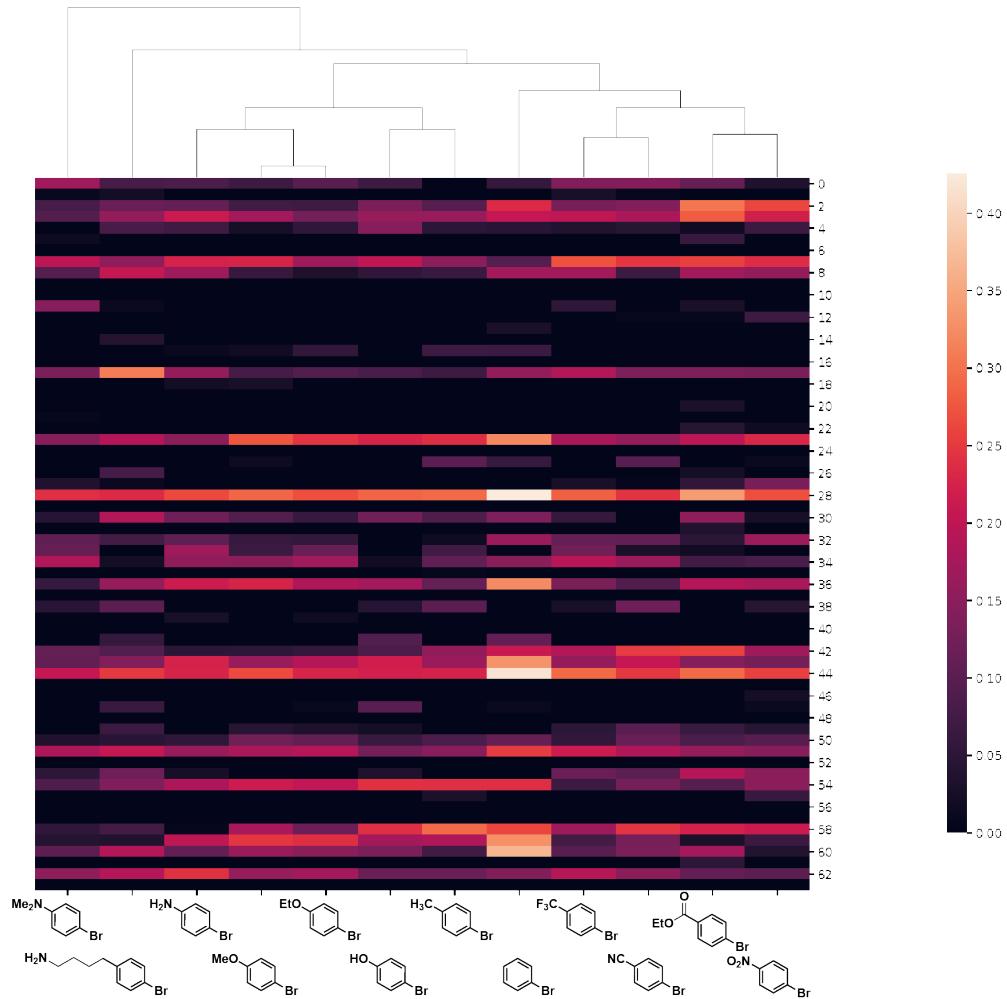
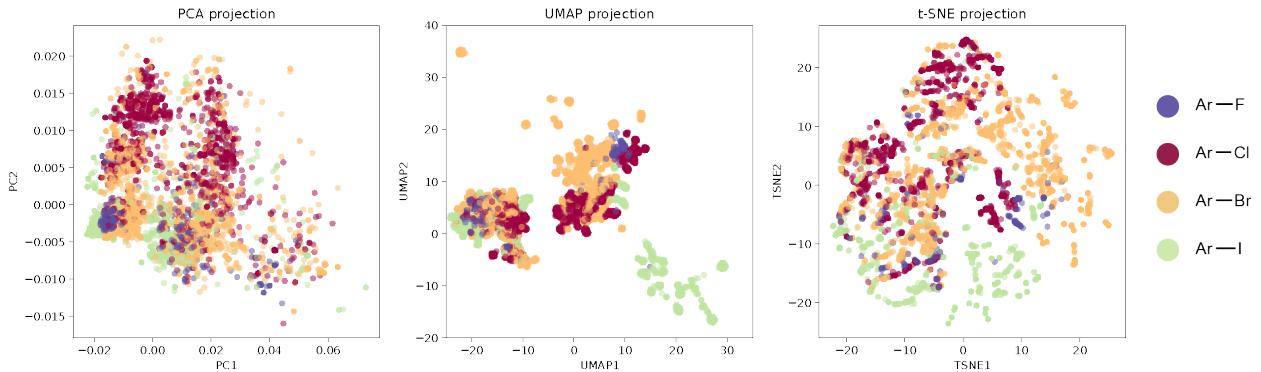


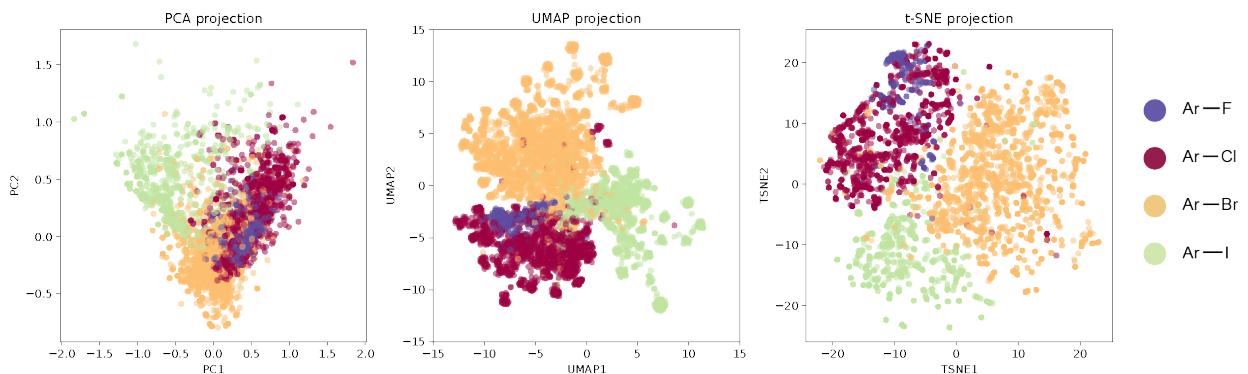
Figure S7: Hierarchical clustering analysis of learned embeddings for a set of aryl bromide molecules. The dendrogram on the top indicates the similarity between the aryl bromide variants. The heatmap below shows the bit value of the corresponding position in the embeddings for each molecule.

Visualization of learned aryl halide chemical spaces

To offer insights into the evolution of the embeddings during training, we present the comparative analysis of our embeddings using PCA, t-SNE, and UMAP projections, both pre- and post-training, in Figure S8. This analysis utilizes the identical set of molecules featured in Figure 2A.



(a) The projection of embeddings before training the GIN using substrate scope groupings.



(b) The projection of embeddings after training the GIN using substrate scope groupings.

Figure S8: The projection of embeddings on a random sampled subset of the substrate scope data. Points are colored by the class of halides.

Correlation with conventional reactivity indicators

Below, we show the analysis of SVM models' regression performance as a function of dataset size on descriptors other than shown in Figure 3C.

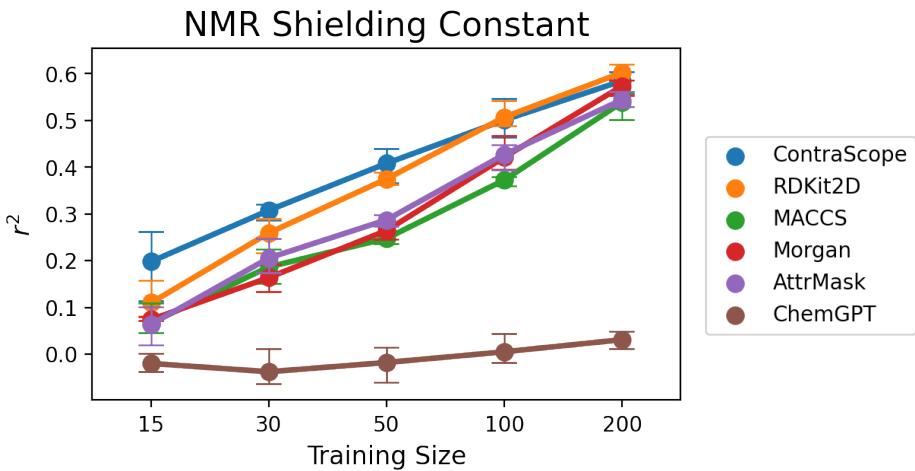


Figure S9: Analysis of SVM regression performance under different dataset size for predicting NMR shielding constants.

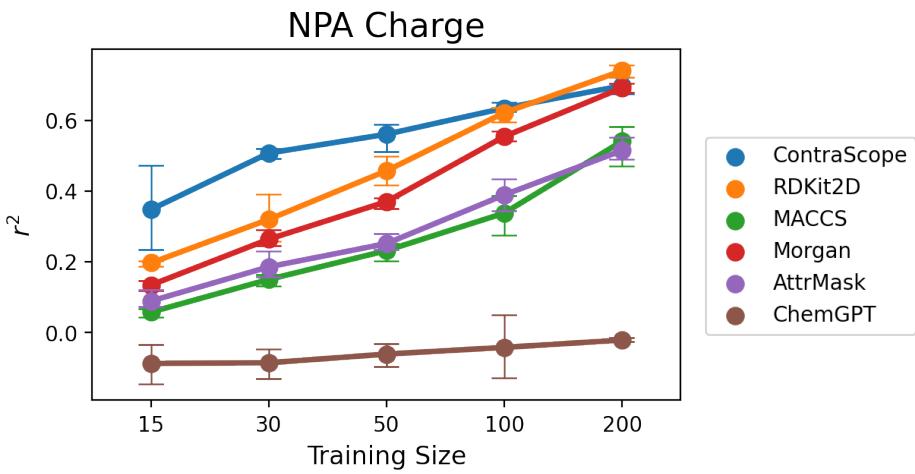


Figure S10: Analysis of SVM regression performance under different dataset size for predicting NPA charges.

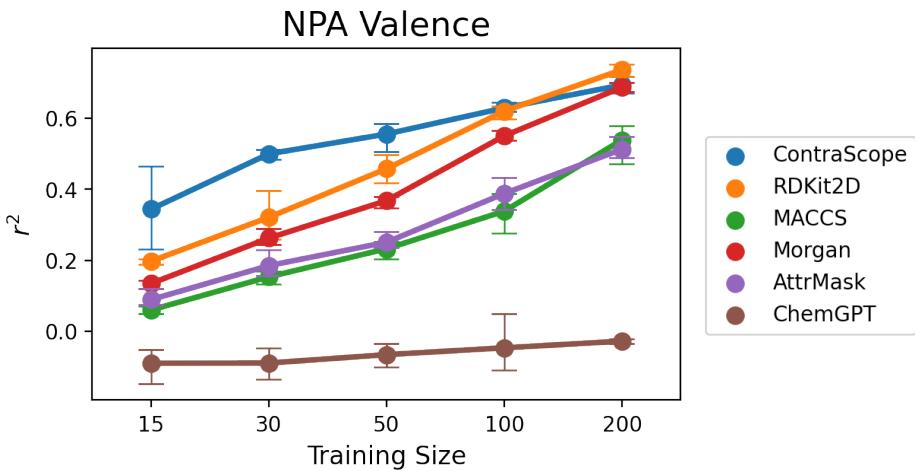


Figure S11: Analysis of SVM regression performance under different dataset size for predicting NPA valences.

Below, we show the t-SNE projection visualizations of the learned embeddings of the 762 aryl halides under the same setting of Figure 2A, colored by traditional reactivity indicator values.

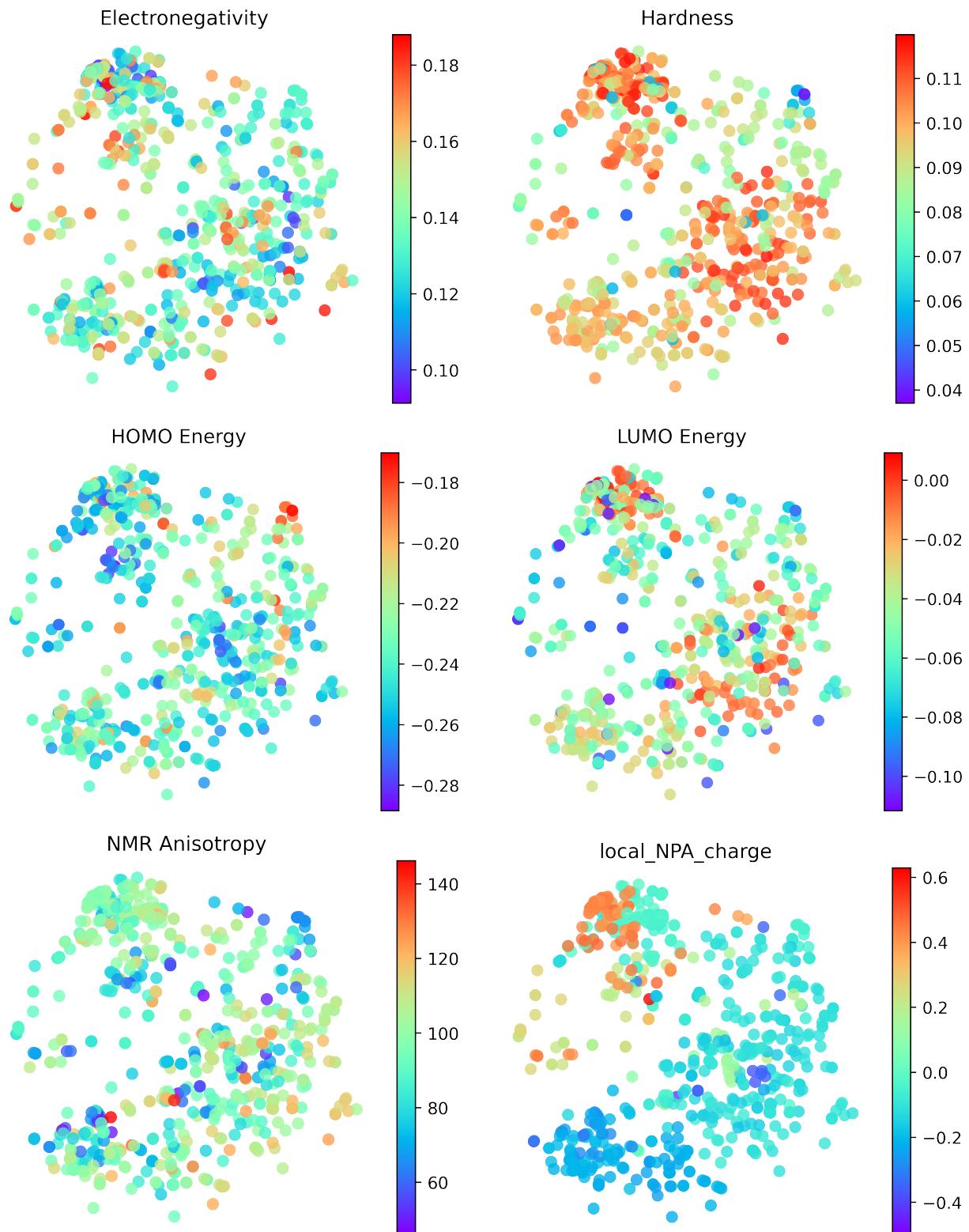


Figure S12: The t-SNE projection colored with various reactivity descriptors.

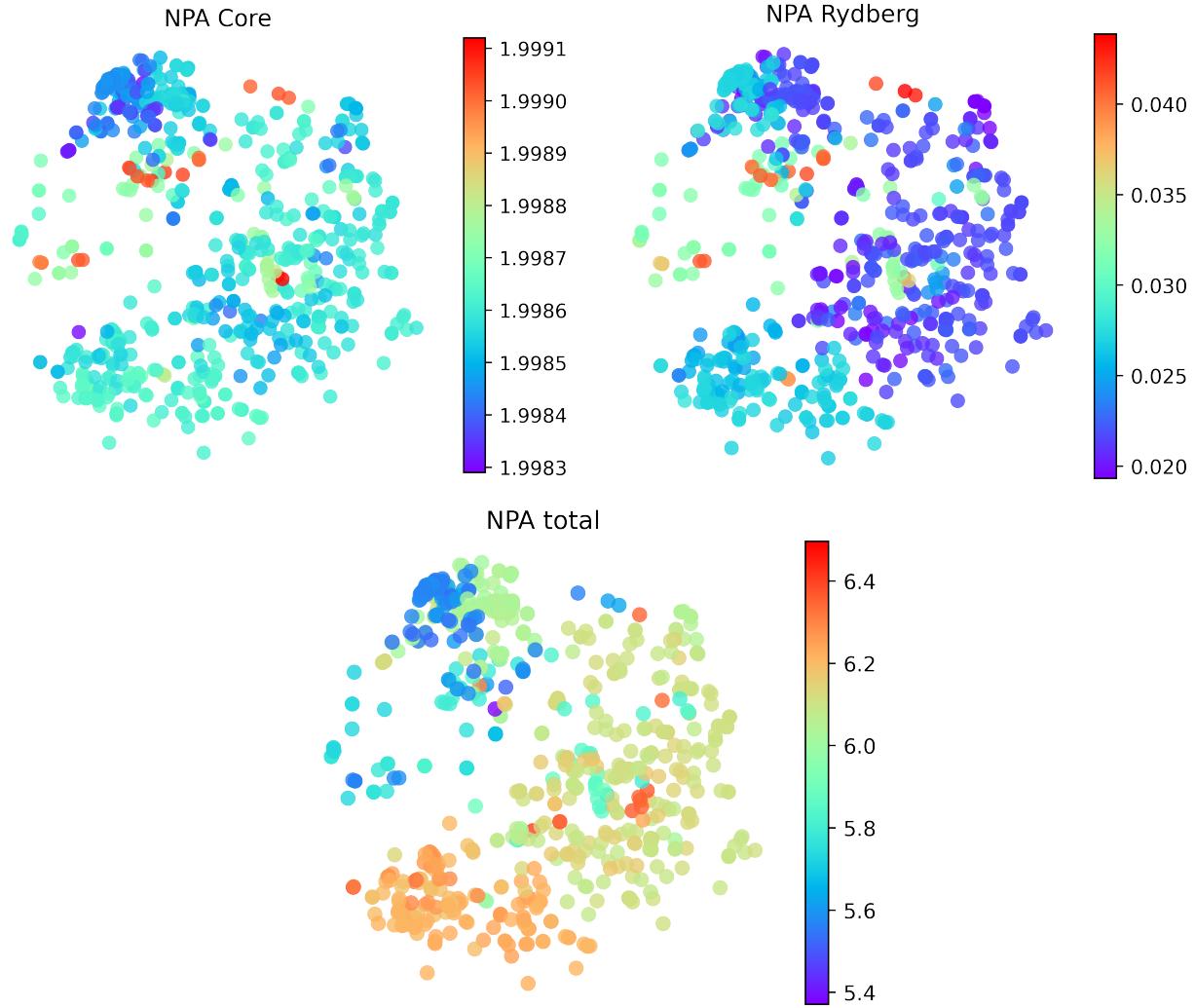


Figure S13: The t-SNE projection colored with various reactivity descriptors. (Continued)

Regression performance on the training substrate scopes

We tried to use the learned embedding as feature to predict reaction yields in the training data. Each substrate scope was treated as a single regression task and leave-one-out validation r^2 for each scope is shown. The methods compared include ContraScope combined with k-Nearest Neighbors (kNN), RDKit2D with kNN, and RDKit2D with Random Forest (RF). The distributions highlight the variability and challenges encountered in the predictive modeling of chemical yields, with all approaches showing a wide distribution of r^2 values, including negative values indicative of a failure to capture any trend.

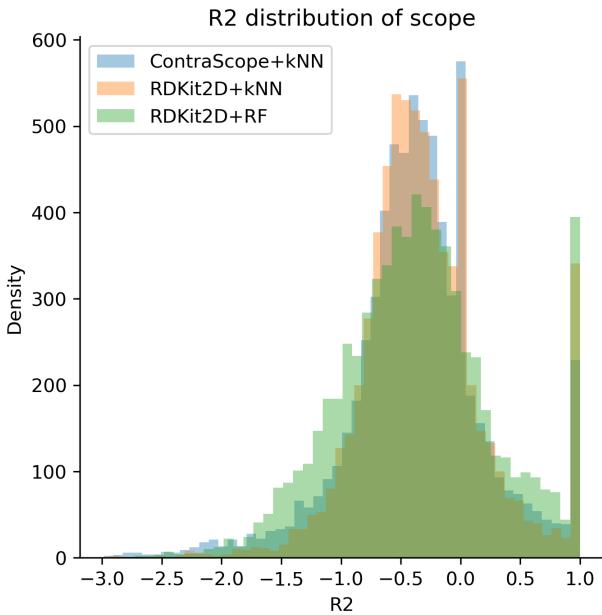


Figure S14: Distribution of the r^2 values for yield prediction across training substrate scopes using nearest neighbor regression.

Investigation on chemistry informers

In the study by Kutchukian et al. (2016),⁶¹ a library of 18 aryl halides was utilized to explore their reactivity in the Buchwald-Hartwig reaction under 18 distinct conditions. We encoded these aryl halides from the specified aryl halide informer library. Subsequently, our analysis involved a comparative assessment of the pair-wise signal-to-noise ratio (SNR) distances derived from the ContraScope embeddings against the pair-wise Euclidean distances computed from the vectors of reported yields. This comparison is graphically represented using heatmaps in Figure S15. However, it is noteworthy that this analysis did not reveal congruent patterns in the left and right panels of the figure.

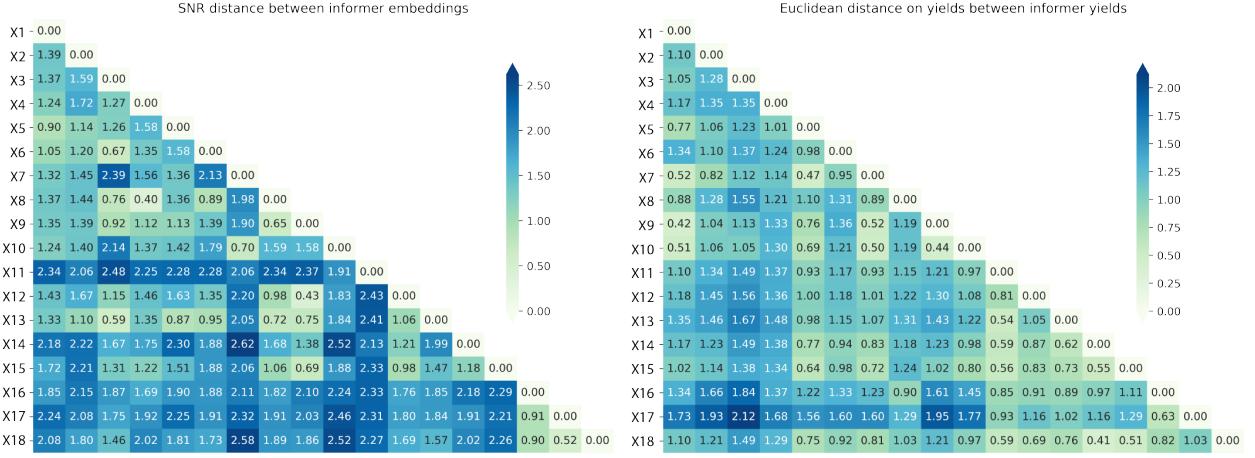


Figure S15: Comparison between the pair-wise SNR distance between the ContraScope embeddings (left), and the pair-wise Euclidean distance between the reported yield vectors (right). Both pair-wise distance matrices are visualized in heatmaps.

Less effective in learning global reactivity contribution

Similar to Figure 3B, we evaluated the regression performance for various molecular embeddings in predicting global chemical reactivity descriptors. A comparison between the performance in global and local descriptors are shown in Figure S16. The results show that while common embeddings yield moderate r^2 values across the descriptors, our embedding exhibits a lower performance, suggesting a less effective capture of global reactivity information in our model.

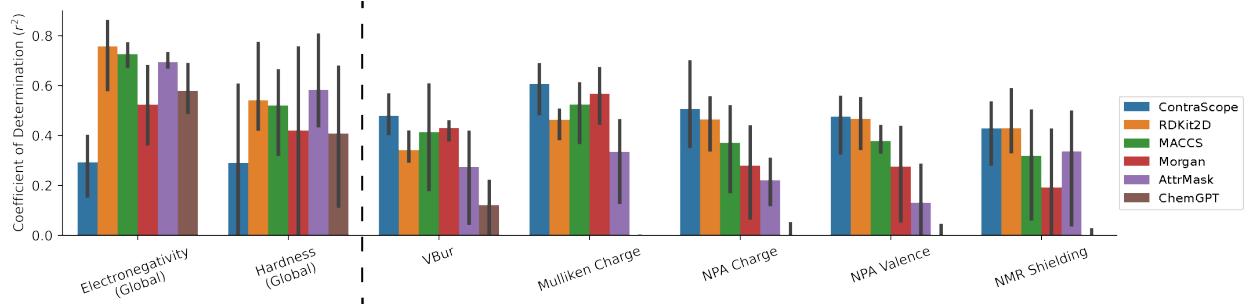


Figure S16: A comparative study of evaluation of regression performance for various molecular embeddings in predicting global and local chemical reactivity descriptors.

Substrate scope comparison

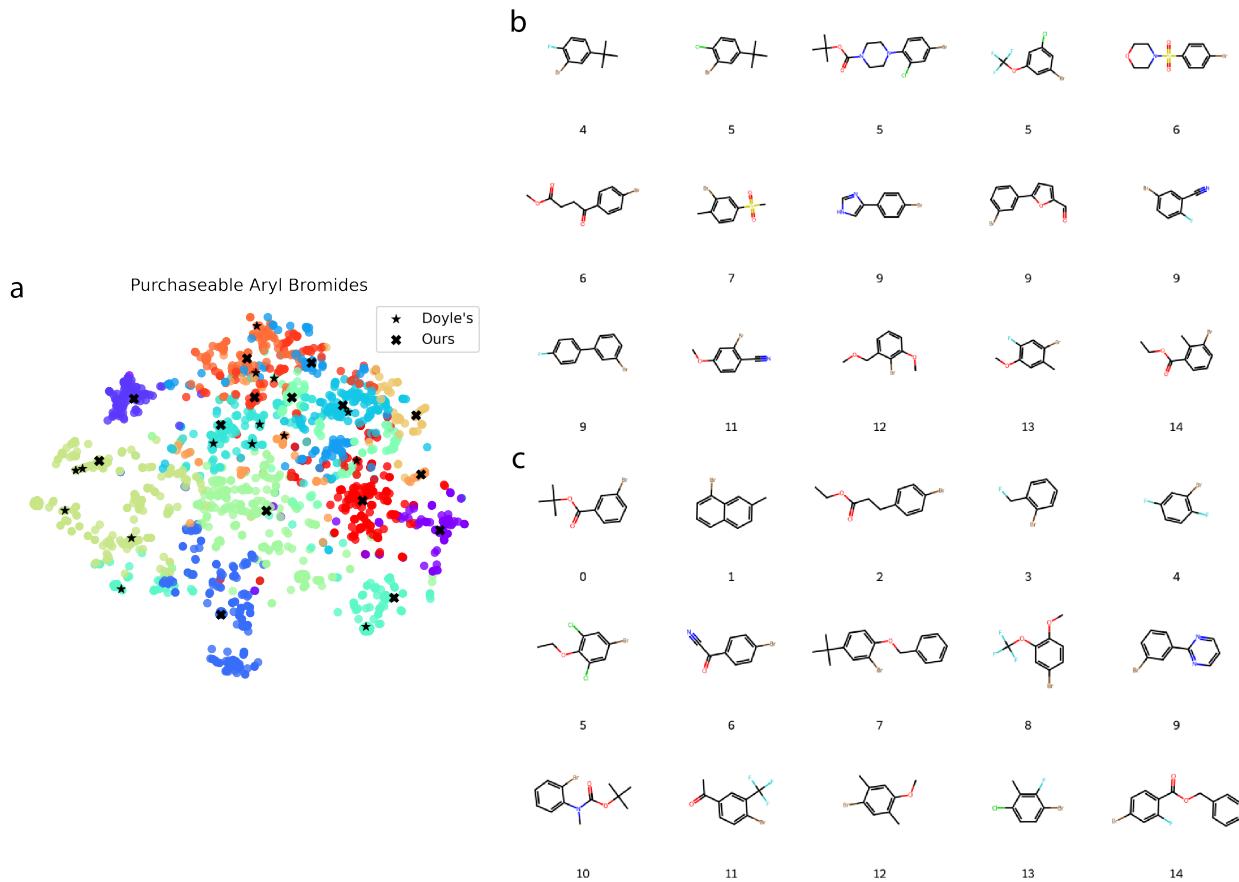


Figure S17: A comparison of substrate scope selected using DFT descriptors and our learned embeddings. (a) a t-SNE projection of all purchasable aryl bromides, categorized based on clustering. Within this projection, substrates selected via DFT descriptors are highlighted with star symbols (labeled as 'Doyle's'), whereas those chosen through our method are marked with cross symbols (denoted as 'Ours'). (b) chemical structures of substrates selected based on DFT descriptors, with accompanying numerical annotations for cluster identification. (c) chemical structures of substrates selected based on our learned embeddings, with accompanying numerical annotations for cluster identification.