6 July 2020

The Photoswitch Dataset: A Molecular Machine Learning Benchmark for the Advancement of Synthetic Chemistry

Aditya Raymond Thawani, Ryan-Rhys Griffiths[1], Arian Jamasb, Anthony Bourached, Penelope Jones, William McCorkindale, Alexander Aldrick, Alpha Lee

1. University of Cambridge

Abstract

The space of synthesizable molecules is greater than 10^{60}, meaning only a vanishingly small fraction of these molecules have ever been realized in the lab. In order to prioritize which regions of this space to explore next, synthetic chemists need access to accurate molecular property predictions. While great advances in molecular machine learning have been made, there is a dearth of benchmarks featuring properties that are useful for the synthetic chemist. Focussing directly on the needs of the synthetic chemist, we introduce the Photoswitch Dataset, a new benchmark for molecular machine learning where improvements in model performance can be immediately observed in the throughput of promising molecules synthesized in the lab. Photoswitches are a versatile class of molecule for medical and renewable energy applications where a molecule's efficacy is governed by its electronic transition wavelengths. We demonstrate superior performance in predicting these wavelengths compared to both time-dependent density functional theory (TD-DFT), the incumbent first principles quantum mechanical approach, as well as a panel of human experts. Our baseline models are currently being deployed in the lab as part of the decision process for candidate synthesis. It is our hope that this benchmark can drive real discoveries in photoswitch chemistry and that future benchmarks can be introduced to pivot learning algorithm development to benefit more expansive areas of synthetic chemistry.

Keywords

Machine Learning, Time-Dependent Density Functional Theory (TDDFT), Big Data, Azoheteroarene, Azopyrazole, Photoswitch, Azobenzene, UV/Vis Spectroscopy, Absorption Spectra, Azophotoswitch

The Photoswitch Dataset: A Molecular Machine Learning Benchmark for the Advancement of Synthetic Chemistry

Aditya R. Thawani*
Department of Chemistry
Imperial College London
London, UK
art12@ic.ac.uk

Ryan-Rhys Griffiths*
Department of Physics
University of Cambridge
Cambridge, UK
rrg27@cam.ac.uk

Arian R. Jamasb
Computer Laboratory
University of Cambridge
Cambridge, UK
arj39@cam.ac.uk

Anthony Bourached
Department of Neurology
University College London
London, UK
ucabab6@ucl.ac.uk

Penelope Jones
Department of Physics
University of Cambridge
Cambridge, UK
pj321@cam.ac.uk

William McCorkindale
Department of Physics
University of Cambridge
Cambridge, UK
wjm41@cam.ac.uk

Alexander Aldrick
Department of Physics
University of Cambridge
Cambridge, UK
av945@cam.ac.uk

Alpha A. Lee
Department of Physics
University of Cambridge
Cambridge, UK
aal44@cam.ac.uk

Abstract

The space of synthesizable molecules is greater than 10^{60}, meaning only a vanishingly small fraction of these molecules have ever been realized in the lab. In order to prioritize which regions of this space to explore next, synthetic chemists need access to accurate molecular property predictions. While great advances in molecular machine learning have been made, there is a dearth of benchmarks featuring properties that are useful for the synthetic chemist. Focussing directly on the needs of the synthetic chemist, we introduce the Photoswitch Dataset, a new benchmark for molecular machine learning where improvements in model performance can be immediately observed in the throughput of promising molecules synthesized in the lab. Photoswitches are a versatile class of molecule for medical and renewable energy applications where a molecule's efficacy is governed by its electronic transition wavelengths. We demonstrate superior performance in predicting these wavelengths compared to both time-dependent density functional theory (TD-DFT), the incumbent first principles quantum mechanical approach, as well as a panel of human experts. Our baseline models are currently being deployed in the lab as part of the decision process for candidate synthesis. It is our hope that this benchmark can drive real discoveries in photoswitch chemistry and that future benchmarks can be introduced to pivot learning algorithm development to benefit more expansive areas of synthetic chemistry.

*Equal contribution

Preprint. Under review.

1 Introduction

In order to prioritize which molecules to synthesize, a chemist would ideally like to know the properties of the molecules in advance. With this information to hand, the amount of time spent on the synthesis of unpromising candidates would be minimized. While there exist first principles computational approaches for property prediction such as time-dependent density functional theory (TD-DFT), these methods are expensive enough to prohibit their use in practice, taking up to 4 days [1] to obtain properties for a single molecule. In light of this, human intuition is still the guiding force behind candidate selection. Advances in molecular machine learning have taken great strides in recent years in areas such as molecule generation [2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20], chemical reaction prediction [21, 22, 23, 24, 25] and molecular property prediction [26, 27, 28, 29, 30]. In particular, accurate machine learning prediction of molecular properties has the potential to cut the attrition rate in the discovery of novel and promising molecules.

Synthetic chemists however, are underserved by current molecular machine learning benchmarks. Research in synthetic chemistry is federated in the sense that a given research field may publish 30 papers containing high-quality experimental data per year. The generalization error of models trained on large benchmark datasets is poor if the molecules contained in the dataset do not reside in the same region of chemical space as the molecules whose properties are being predicted. In collaboration with synthetic chemists, we aim to address this problem through the introduction of benchmarks where the domain of the dataset is matched to the search space of interest. In this paper we introduce one such benchmark, the Photoswitch Dataset, to enable accurate predictions for the transition wavelength properties of photoswitch molecules, depicted in Figure 1.

Our goals for the benchmark may be articulated as follows:

1. Perform faster prediction relative to TD-DFT.
2. Obtain improved accuracy relative to human experts.
3. Operationalize model predictions in the context of laboratory synthesis.

We achieve all of these goals with baseline models and we hope that further progress can be made by users of the benchmark towards achieving an absolute error of less than 10 nm for all molecules of interest, a threshold that would reduce the number of poor performing molecules in this field to zero. We suggest how best this might be accomplished through the development of models that leverage out-of-domain data to improve performance, models that account for solvent effects in the observed wavelength value as well as models where prediction error may be attributed to aspects of the molecular representation.

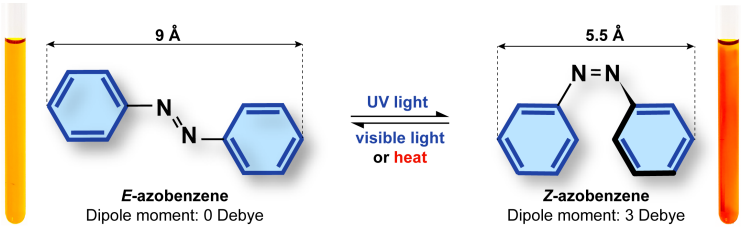


Figure 1: Photoswitch molecules can be reversibly converted between multiple structural states using light. Azobenzene is an example of a photoswitchable molecule; analogues of which are referred to as azophotoswitches. The two isomers of azobenzene have different conformations and physicochemical properties. Photopharmacology leverages these dynamic properties in light-addressable drugs; where the inactive isomer may be converted *in situ* to the biologically active isomer [31, 32, 33, 34]. Solar thermal energy fuels use azophotoswitches to capture energy from the sun and store it for later release [35, 36, 37].

2 Related Work

Arguably one of the most well-known molecular machine learning benchmarks is MoleculeNet [28], a dataset collection comprising experimentally-determined properties including hydration free energies

[38] and aqueous solubilities [39], in addition to computed properties such as atomization energies [40, 41, 42], excitation energies [43, 44] geometric, energetic, electronic, and thermodynamic properties [43, 45]. In the graph machine learning community benchmark datasets feature properties such as mutagenicity [46] as well as synthetically useful properties such as binding affinities [47, 48]. To our knowledge this is the first benchmark featuring experimentally-determined transition wavelengths for photoswitch molecules. Our dataset has been manually-curated by an expert photoswitch chemist and represents a vital extension of the current benchmark ecosystem to incorporate the properties of light-activated molecules.

In the context of transition wavelength prediction there has been prior work on learning corrections to TD-DFT [44] as well as directly predicting experimental values [49]. In terms of open-source data on photoswitch molecules, ChEBI [50], a notable chembioinformatics database, currently lists ca. 40 biologically active azophotoswitches of minimalistic structural diversity whilst PubChem [51, 52] possesses over 3900 azophotoswitch structures with no associated photophysical data. Recently, Cole and co-workers [53] reported a dataset of 18000 individual molecules with photophysical data. Whilst ca. 500 of these are azophotoswitches none of these scaffolds cover the latest generations of azophotoswitches currently deployed in the laboratory. Additionally, the photophysical data provided does not include information on the *Z* isomer.

3 An Overview of the Photoswitch Dataset

3.1 Data Acquisition

We engage a trained photoswitch chemist to source experimentally-determined properties of photoswitch molecules reported in chemistry journals. To ensure molecular diversity, we include azobenzene derivatives with a diverse range of substitution patterns and functional groups. This is vitally important from a synthetic point-of-view as such functional groups serve as handles for further synthetic modification. Furthermore, we also include the latest generation of azoheteroarenes and cyclic azobenzenes which have established themselves as possessing superior photophysical properties to unmodified azobenzene. In all instances, we extract the following photophysical properties for each molecule, where available: rate of thermal isomerization, photostationary state, experimental and TD-DFT-computed $\pi - \pi^*$ and $n - \pi^*$ transition wavelengths, the molar extinction coefficient, Wiberg index, irradiation wavelength in a and irradiation solvent. A full list of references for the data sources is provided in Appendix A in addition to background information on the tabulated properties.

3.2 Tasks

We focus our experimentally-determined property prediction benchmark on the following 4 tasks:

1. ***E* isomer $\pi - \pi^*$ Transition wavelength prediction**: 392 molecules
2. ***E* isomer $n-\pi^*$ Transition wavelength prediction**: 141 molecules
3. ***Z* isomer $\pi - \pi^*$ Transition wavelength prediction**: 93 molecules
4. ***Z* isomer $n-\pi^*$ Transition wavelength prediction**: 123 molecules

Where the dataset size is given for each task. The aforementioned properties are of particular importance because they are the core determiners of quantitative, bidirectional photoswitching. As an illustrative example, a photoswitchable drug works on the premise that one isomer is biologically active and has a therapeutic effect, whilst the other isomer does not. Ideally, the inactive isomer is dosed to the patient and converted using light to the active isomer. A drug capable of quantitative photoswitching is thus more potent than an incomplete photoswitching analogue and could be administered at a lower dose which minimizes side effects without compromising on therapeutic potential. This effect is best achieved when molecules possess band separation within their absorption spectra such that a wavelength of light can be chosen to attain selective and complete photoisomerisation. Accurately predicting the $\pi - \pi^*$ and $n - \pi^*$ transition wavelengths for both sets of isomers is critical to determining if this property can be fulfilled. The experimental setup for these tasks is described in subsection 4.1.

3.3 How to Use the Benchmark

We release our benchmark repository at the following link and will maintain a public leaderboard for each of the 3 tasks. Of particular interest are submissions which:

1. Achieve good performance on the prediction tasks by leveraging data that is not currently contained in the repository. Such results may point either towards further sources of in-domain data relevant for the prediction task or to methods that manage to achieve good generalization performance using out-of-domain data.
2. Graph-based methods that can account for solvent effects on the transition wavelength via a FiLM mechanism [54] or by incorporating additional node features. Solvent identity is provided for all measurements.
3. Models where prediction error is interpretable and may be attributed to characteristics of the molecular representation.

3.4 Dataset Visualization

We provide visualizations of the dataset under different molecular representations using the UMAP algorithm [55] in Appendix B.

4 Experiments

In this section we evaluate a diverse set of models and molecular representations on the 4 wavelength prediction tasks constituting the property prediction benchmark. Through an analysis of the model prediction errors we are able to identify missing inductive biases in the molecular representation, leading us to propose a new hybrid representation which achieves the best results on the first benchmark task. We take this representation forward to an independent evaluation against the predictions of TD-DFT as well as a panel of synthetic photoswitch chemists. Lastly, we demonstrate the need for an expert-in-the-loop when designing property prediction benchmarks tailored to real-world synthesis. We show that a model trained on a large dataset of 6142 chemically distinct molecules does not improve performance on our benchmark. A large, out-of-distribution dataset is less useful than a small in-distribution one. Full details of all experiments, in addition to method background may be found in Appendix C.

4.1 Wavelength Prediction Benchmark

We evaluate performance on 20 random train/test splits in a ratio of 80/20 using the root mean square error (RMSE), mean absolute error (MAE) and coefficient of determination (R^2) as performance metrics, reporting the mean and standard error for each metric. We evaluate the following models: Random Forest (RF), Gaussian Processes (GP), Attentive Neural Processes (ANP) [56], Graph Convolutional Networks (GCN) [57], Graph Attention Networks (GAT) [58], Directed Message-Passing Neural Networks (DMPNN) [59] and the following representations: Morgan fingerprints [60], RDKit fragments [61] and smooth overlap of atomic positions (SOAP) [62]. In addition we introduce a hybrid representation, fragprints, the motivation for which is outlined in subsection 4.2. For the purpose of the benchmark, hyperparameter selection for GP-based approaches is performed by optimizing the marginal likelihood on the train set whereas for other methods cross-validation may be performed.

Random Forest is trained using scikit-learn [63] with 1000 estimators and a maximum depth of 300. We implement a Gaussian Process in GPflow [64] using a Tanimoto kernel [65]. We set the mean function to be the empirical mean of the data and treat the kernel variance and likelihood variance as hyperparameters, optimizing their values under the marginal likelihood. For the attentive neural process we use 2 hidden layers of dimension 32 for each of the decoder, latent decoder and the deterministic encoder respectively, 8-dimensional latent variables r and z, and run 500 iterations with the Adam optimizer [66] with a learning rate of 0.001. For the ANP we perform principal components regression by reducing the representation dimension to 50. We implement Graph Convolutional Networks and Graph Attention Networks in the DGL-LifeSci library [67]. Node features include one-hot representations of atom-type, atom degree, the number of implicit hydrogen atoms attached

to each atom, the total number of hydrogen atoms per atom, atom hybridization, the formal charge and number of radical electrons on the atom. Edge features contain one-hot encodings of bond-type and Booleans indicating the stereogenic configuration of the bond and whether the bond is conjugated or in a ring. For the GCN we use two hidden layers with 32 hidden units and ReLU activations, applying BatchNorm [68] to both layers. The remaining parameters are the default library values. For the GAT we use two hidden layers with 32 units each, 4 attention heads, an alpha value of 0.2 in both layers and ELU activations, using the default library values for the remaining parameters. We use a single Directed Message-Passing Neural Network model trained for 50 epochs, with additional normalized 2D RDKit features and without performing Bayesian optimization over the parameters defining the model architecture. All remaining parameters were set to the default values in [59].

For representations, we use 2048-bit Morgan fingerprints with a bond radius of 3 implemented in RDKit [61]. We use 85-dimensional fragment features computed using the RDKit descriptors module. We use the Dscribe library [69] to compute SOAP descriptors using an `rcut` parameter of 3.0 a `sigma` value of 0.2. An `nmax` parameter of 12 and an `lmax` parameter of 8. We use an REMatch kernel with polynomial base kernel of degree 3.0, `gamma` = 1.0, `coef0` = 0, `alpha` = 0.5 and `threshold` = $1e^{-6}$. We apply standardization to the property values in all experiments. The results of the aforementioned models and representations are given in Table 6. Additional results including Message-passing neural networks (MPNN) [70], a black-box alpha divergence minimization Bayesian neural network (BNN) [71] and an LSTM with augmented SMILES, SMILES-X [72] are presented in appendix C.

We note that featurizations using standard molecular descriptors are more than competitive with neural representations for this dataset. The best-performing representation/model pair was the GP-Tanimoto kernel and our own hybrid descriptor set "fragprints". We describe the rationale for developing this representation in the following section.

4.2 Prediction Error as a Guide to Representation Selection

On the *E* isomer $\pi - \pi^*$ transition wavelength prediction task, we note occasionally marked discrepancies in the predictions made under the Morgan fingerprint and fragment representations. We show one such discrepancy in Figure 2. The resultant analysis motivated the expansion of the molecular feature set to include both representations as "fragprints"

4.3 TD-DFT Comparison

We compare the top-performing model from the benchmark, the Gaussian Process, Tanimoto kernel and fragprints combination against two widely-utilized levels of time-dependent density functional theory: CAM-B3LYP [73] and PBE0 [74, 75]. While the CAM-B3LYP level of theory offers highly accurate predictions, its computational cost is a stumbling block towards its adoption in synthetic photoswitch chemistry. To obtain the predictions for a single photoswitch molecule one is required to perform a one-day ground state energy minimization followed by a one-day TD-DFT calculation [1]. In the photoswitch chemist's case these calculations need to be performed for both molecular isomers leading to a wall-clock time of 4 days in total per molecule. When screening multiple molecules is desirable this cost is prohibitive and so in practice it is easier to screen candidates based on human chemical intuition. In contrast, inference in a data-driven model is in on the order of seconds.

In Table 2 and Table 3 we present the performance comparison against 99 molecules and 114 molecules for CAM-B3LYP and PBE0 respectively taken from the results of a benchmark quantum chemistry study [76]. For the GP model, we perform leave-one-out validation, testing on a single molecule and training on the others in addition to the experimentally-determined property values for molecules acquired from synthesis papers. We then average the predictions errors and report the standard error. The model outperforms PBE0 by a large margin and provides comparable performance to CAM-B3LYP. Further background on time-dependent density functional theory are available in Appendix D.

tesubsectionHuman Performance Comparison

In practice, candidate screening is undertaken based on the opinion of a human expert due to the speed at which predictions may be obtained. While inference in a data-driven model is comparable to the human approach in terms of speed, we aim in this section to compare the predictive accuracy of the two approaches. In order to achieve this, we assemble a panel of 14 human experts, comprising

Table 1: Test Set Performance in Predicting the Transition Wavelengths of the E and Z isomers. Best-performing models are highlighted in bold.

	E isomer $\pi - \pi^*$ (nm)	*E* isomer $n-\pi^*$ (nm)	Z isomer $\pi - \pi^*$ (nm)	*Z* isomer $n-\pi^*$ (nm)
RMSE				
RF + Morgan	25.3 ± 0.9	$\mathbf{10.2 \pm 0.4}$	14.0 ± 0.6	11.1 ± 0.4
RF + Fragments	26.4 ± 1.1	11.4 ± 0.5	17.0 ± 0.8	14.2 ± 0.6
RF + Fragprints	23.4 ± 0.9	11.0 ± 0.4	14.2 ± 0.6	11.3 ± 0.6
GP + Morgan	23.4 ± 0.8	11.4 ± 0.5	13.2 ± 0.7	$\mathbf{11.0 \pm 0.7}$
GP + Fragments	26.3 ± 0.8	11.6 ± 0.5	15.5 ± 0.8	12.6 ± 0.5
GP + Fragprints	$\mathbf{20.9 \pm 0.7}$	11.1 ± 0.5	$\mathbf{13.1 \pm 0.6}$	11.4 ± 0.7
GP + SOAP	21.0 ± 0.6	22.7 ± 0.6	17.8 ± 0.8	15.0 ± 0.5
ANP + Morgan	28.1 ± 1.3	13.6 ± 0.5	13.5 ± 0.6	$\mathbf{11.0 \pm 0.6}$
ANP + Fragments	27.9 ± 1.1	13.8 ± 0.9	17.2 ± 0.8	14.1 ± 0.7
ANP + Fragprints	27.0 ± 0.8	11.6 ± 0.5	14.5 ± 0.8	11.3 ± 0.7
GCN	22.0 ± 0.8	12.8 ± 0.8	16.3 ± 0.8	13.1 ± 0.8
GAT	26.4 ± 1.1	16.9 ± 1.9	19.6 ± 1.0	14.5 ± 0.8
DMPNN	27.1 ± 1.4	13.9 ± 0.6	17.5 ± 0.7	13.8 ± 0.4
MAE				
RF + Morgan	15.5 ± 0.5	$\mathbf{7.3 \pm 0.3}$	10.1 ± 0.4	$\mathbf{6.6 \pm 0.3}$
RF + Fragments	16.4 ± 0.5	8.5 ± 0.3	12.2 ± 0.6	9.0 ± 0.4
RF + Fragprints	13.9 ± 0.4	7.7 ± 0.3	10.0 ± 0.4	6.8 ± 0.3
GP + Morgan	15.2 ± 0.4	8.4 ± 0.3	9.8 ± 0.4	6.9 ± 0.3
GP + Fragments	17.3 ± 0.4	8.6 ± 0.3	11.5 ± 0.5	8.2 ± 0.3
GP + Fragprints	$\mathbf{13.3 \pm 0.3}$	8.2 ± 0.3	$\mathbf{9.8 \pm 0.4}$	7.1 ± 0.3
GP + SOAP	14.3 ± 0.3	19.3 ± 0.5	12.9 ± 0.6	11.4 ± 0.4
ANP + Morgan	17.9 ± 0.7	10.1 ± 0.4	10.0 ± 0.4	7.2 ± 0.3
ANP + Fragments	17.4 ± 0.6	9.4 ± 0.4	12.3 ± 0.6	8.9 ± 0.4
ANP + Fragprints	18.1 ± 0.5	8.6 ± 0.3	10.4 ± 0.5	7.0 ± 0.3
GCN	13.9 ± 0.3	8.6 ± 0.3	11.6 ± 0.5	8.6 ± 0.5
GAT	18.1 ± 0.7	10.7 ± 0.6	14.4 ± 0.8	10.8 ± 0.7
DMPNN	17.1 ± 0.8	10.6 ± 0.4	12.8 ± 0.6	9.8 ± 0.3
$\boldsymbol{R^2}$				
RF + Morgan	0.85 ± 0.01	$\mathbf{0.80 \pm 0.01}$	0.25 ± 0.06	0.36 ± 0.06
RF + Fragments	0.83 ± 0.01	0.75 ± 0.02	-0.15 ± 0.11	-0.05 ± 0.07
RF + Fragprints	0.87 ± 0.01	0.77 ± 0.02	0.23 ± 0.07	0.33 ± 0.06
GP + Morgan	0.87 ± 0.01	0.76 ± 0.01	0.34 ± 0.05	$\mathbf{0.38 \pm 0.05}$
GP + Fragments	0.84 ± 0.01	0.74 ± 0.02	0.07 ± 0.08	0.19 ± 0.05
GP + Fragprints	$\mathbf{0.90 \pm 0.01}$	0.77 ± 0.02	$\mathbf{0.35 \pm 0.05}$	0.33 ± 0.05
GP + SOAP	0.89 ± 0.01	-0.08 ± 0.03	-0.05 ± 0.02	-0.07 ± 0.02
ANP + Morgan	0.70 ± 0.02	0.66 ± 0.02	0.30 ± 0.06	$\mathbf{0.38 \pm 0.05}$
ANP + Fragments	0.81 ± 0.01	0.62 ± 0.05	-0.16 ± 0.11	-0.06 ± 0.10
ANP + Fragprints	0.83 ± 0.01	0.75 ± 0.01	0.18 ± 0.08	0.35 ± 0.05
GCN	0.87 ± 0.01	0.66 ± 0.03	-0.41 ± 0.22	-0.92 ± 0.3
GAT	0.81 ± 0.02	0.57 ± 0.04	0.39 ± 0.17	-1.07 ± 0.4
DMPNN	0.82 ± 0.02	0.63 ± 0.02	-0.05 ± 0.07	0.11 ± 0.04

Table 2: TD-DFT Benchmark Comparison against 99 Molecules at the CAM-B3LYP Level of Theory.

Method	MAE (nm)
GP + Fragprints	$\mathbf{14.9 \pm 1.4}$
CAM-B3LYP	16.5 ± 1.6

Table 3: TD-DFT Benchmark Comparison against 114 Molecules at the PBE0 Level of Theory.

Method	MAE (nm)
GP + Fragprints	$\mathbf{15.2 \pm 1.3}$
PBE0	26.0 ± 1.8

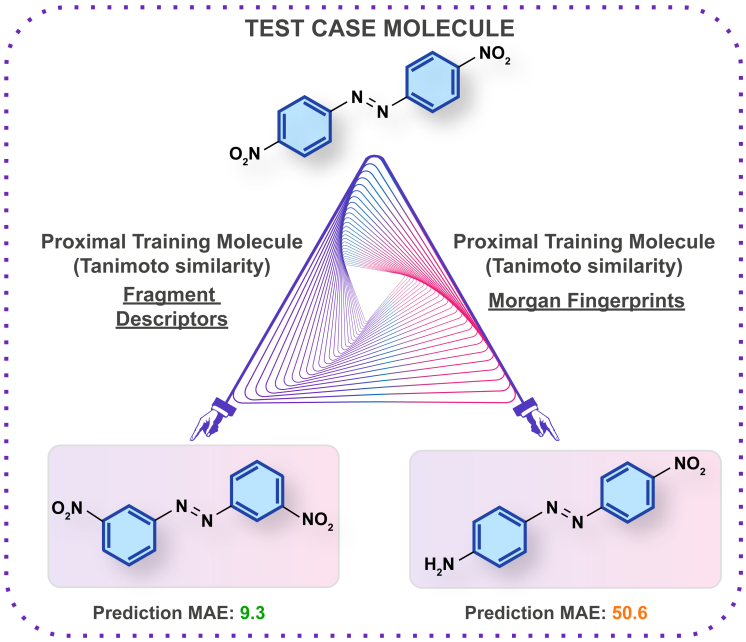


Figure 2: An analysis of the prediction errors under the Morgan fingerprint and fragment representations. The molecule on which the prediction is being made is located at the apex of the triangle with the proximal training molecule at the base. Fragment descriptors identify another di-substituted nitroazobenzene as the most similar molecule contained in the train set. By contrast, Morgan fingerprints identify a molecule in possession of a similar substitution pattern to the test case, but with different functionalization. On this particular test instance it is the identity of the functional groups rather than the substitution pattern which dictates the wavelength properties and hence fragment descriptors achieve a much lower error. As such, although fingerprints offer better overall performance, fragments are clearly informative features for certain test cases.

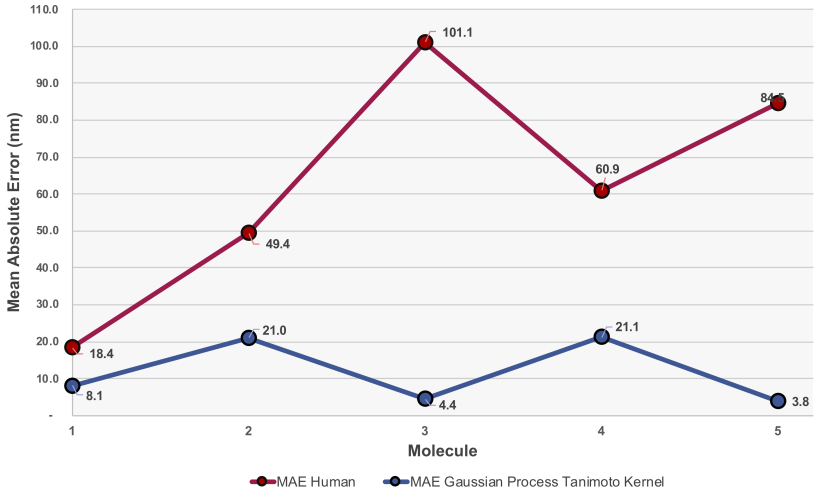


Figure 3: A performance comparison between human experts and the GP-fragprints model. MAEs are computed on a per molecule basis across all human participants.

Table 4: Generalization Performance of RF-fragprints trained on an out-of-domain dataset of 6142 molecules and evaluated on the Photoswitch Dataset.

RMSE	85.2
MAE	72.5
$\boldsymbol{R}^2$	-0.66

Table 5: Generalization Performance of RF-fragprints on the original *E* isomer $\pi - \pi^*$ transition wavelength prediction task where the train set has been augmented with the out-of-domain dataset of 6142 molecules.

RMSE	36.9 ± 1.2
MAE	22.7 ± 0.7
$\boldsymbol{R}^2$	0.67 ± 0.02

Postdoctoral Research Assistants and PhD students in synthetic chemistry with a combined research experience of $>$100 years. The assigned task is to predict the *E* isomer $\pi - \pi^*$ transition wavelength for five molecules taken from the dataset. A reference molecule is also provided with associated $\pi - \pi^*$ wavelength. The reference molecule possesses either single, double or triple point changes from the target molecule and serves to mimic the laboratory decision-making process of predicting an unknown molecule's property with respect to a known one. In all instances, those polled have received formal training in the fundamentals of UV-Vis spectroscopy.

Analysing the MAE across all humans per molecule (Figure 3), we note that the humans perform worse than the GP-fragprints model in all instances. In going from molecule 1 to 5, the number of point changes on the molecule increases steadily, thus, increasing the difficulty of prediction. Noticeably, the human performance is approximately five-fold worse on molecule 5 (three point changes) relative to molecule 1 (one point change).

This highlights the fact that in instances of multiple functional group modifications, human experts are unable to reliably predict the impact on the *E* isomer $\pi - \pi^*$ transition wavelength. Thus, a computational aid with good predictive performance and understanding of the underlying functional group patterns and trends is sorely needed to assist the human decision making process as to which molecules to pursue in the synthetic laboratory.

4.4 Out-of-Domain Generalization

In this section we evaluate the generalization performance of a model trained on the *E* isomer $\pi - \pi^*$ values of a large dataset of 6142 out-of-domain molecules from [53]. We train a Random Forest regressor implemented in the scikit-learn library with 1000 estimators and a max depth of 300 on the fragprint representation of the molecules. In Table 4 we present results for the case when the train set consists of the large dataset of 6142 molecules and the test set consists of the entire photoswitch dataset. In Table 5 we present the results on the original *E* isomer $\pi - \pi^*$ transition wavelength prediction task from subsection 4.1 where the train set of each random 80/20 train/test split is augmented with the molecules from the large dataset. The results indicate that the data for out-of-domain molecules provides no benefit for the prediction task and even degrades performance relative to training on in-domain data only.

Based on these results we highlight the importance of designing synthetic molecular machine learning benchmarks with a real-world application in mind and where possible, involving synthetic chemists in the curation process. By targeted data collation on a narrow and well-defined region of chemical space where the molecules are in-domain relative to the task, we may mitigate generalization error. We leave it as an open problem for users to demonstrate superior predictive performance on the benchmark tasks using data that is not currently present in the Photoswitch Dataset repository.

5 Conclusions

We have introduced the Photoswitch Dataset, a molecular machine learning benchmark to accelerate the discovery of promising light-activated molecules. We highlight the utility of the benchmark to enable the direct operationalization of model predictions in the lab by exhibiting baseline models capable of:

1. Comparable prediction accuracy to TD-DFT at a fraction of the time.
2. Comparable prediction time to human experts with much higher accuracy.

Through maintaining a public leaderboard on the benchmark tasks, we seek to reduce prediction error to ca. 10 nm for all molecules of interest. At this level of accuracy it would become possible to eliminate failed synthesis attempts at a cost of 2-3 weeks per molecule. Of broader interest for molecular machine learning are model submissions capable of achieving high generalization performance without making use of data from the Photoswitch Dataset, models which can successfully account for solvent effects in the molecular representation or otherwise as well as models where prediction error may be attributable to the facets of a particular molecular representation. Lastly, we hope that future molecular machine learning benchmarks may be developed in consultation with synthetic chemists, forming a pipeline that supplies model predictions directly to the end-user.

6 Author Contributions

A.R.T and R-R.G conceived the project. A.A.L proposed that the benchmark dataset be constructed by a domain expert. A.R.T curated the experimental data from the literature. A.R.T and R-R.G jointly devised all experiments. A.B proposed visualizing the dataset using the UMAP algorithm. A.J implemented all graph-based models save for the DMPNN. P.J implemented the Attentive Neural Process. W.M implemented the GP-SOAP model as well as the DMPNN. A.A provided context in the design and interpretation of the DFT comparison and wrote the background section on TD-DFT. R-R.G and A.R.T wrote the manuscript. R-R.G implemented all remaining algorithms and maintains the codebase.

7 Acknowledgements

A.R.T would like to express his sincere gratitude to Professor Matthew J. Fuchter and associated research group members for support rendered and participation in the human performance comparison.

8 Competing Interests

The authors declare that they have no competing financial interests.

References

[1] Anatoly M Belostotskii. *Conformational Concept for Synthetic Chemist's Use*. World Scientific, 2015. doi: 10.1142/6832. URL https://www.worldscientific.com/doi/abs/10.1142/6832.

[2] Rafael Gómez-Bombarelli, Jennifer N Wei, David Duvenaud, José Miguel Hernández-Lobato, Benjamín Sánchez-Lengeling, Dennis Sheberla, Jorge Aguilera-Iparraguirre, Timothy D Hirzel, Ryan P Adams, and Alán Aspuru-Guzik. Automatic chemical design using a data-driven continuous representation of molecules. *ACS Central Science*, 4(2):268–276, 2018.

[3] Jaechang Lim, Sang-Yeon Hwang, Seokhyun Moon, Seungsu Kim, and Woo Youn Kim. Scaffold-based molecular design with a graph generative model. *Chem. Sci.*, 11:1153–1164, 2020. doi: 10.1039/C9SC04503A.

[4] Matt J Kusner, Brooks Paige, and José Miguel Hernández-Lobato. Grammar variational autoencoder. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pages 1945–1954, 2017.

[5] Wengong Jin, Regina Barzilay, and Tommi Jaakkola. Junction tree variational autoencoder for molecular graph generation. In *International Conference on Machine Learning*, pages 2323–2332, 2018.

[6] Jaechang Lim, Seongok Ryu, Jin Woo Kim, and Woo Youn Kim. Molecular generative model based on conditional variational autoencoder for de novo molecular design. *Journal of cheminformatics*, 10(1):1–9, 2018.

[7] Ryan-Rhys Griffiths and José Miguel Hernández-Lobato. Constrained Bayesian optimization for automatic chemical design using variational autoencoders. *Chemical Science*, 2020.

[8] Sai Krishna Gottipati, Boris Sattarov, Sufeng Niu, Yashaswi Pathak, Haoran Wei, Shengchao Liu, Karam MJ Thomas, Simon Blackburn, Connor W Coley, Jian Tang, et al. Learning to navigate the synthetically accessible chemical space using reinforcement learning. *arXiv preprint arXiv:2004.12485*, 2020.

[9] Seung Hwan Hong, Seongok Ryu, Jaechang Lim, and Woo Youn Kim. Molecular generative model based on adversarially regularized autoencoder. *Journal of Chemical Information and Modeling*, 2019.

[10] Daniel C Elton, Zois Boukouvalas, Mark D Fuge, and Peter W Chung. Deep learning for molecular design—a review of the state of the art. *Molecular Systems Design & Engineering*, 4(4):828–849, 2019.

[11] Daniil Polykovskiy, Alexander Zhebrak, Benjamin Sanchez-Lengeling, Sergey Golovanov, Oktai Tatanov, Stanislav Belyaev, Rauf Kurbanov, Aleksey Artamonov, Vladimir Aladinskiy, Mark Veselov, et al. Molecular sets (moses): a benchmarking platform for molecular generation models. *arXiv preprint arXiv:1811.12823*, 2018.

[12] Nathan Brown, Marco Fiscato, Marwin HS Segler, and Alain C Vaucher. Guacamol: benchmarking models for de novo molecular design. *Journal of chemical information and modeling*, 59(3):1096–1108, 2019.

[13] Omar Mahmood and José Miguel Hernández-Lobato. A cold approach to generating optimal samples. *arXiv preprint arXiv:1905.09885*, 2019.

[14] Boris Sattarov, Igor I Baskin, Dragos Horvath, Gilles Marcou, Esben Jannik Bjerrum, and Alexandre Varnek. De novo molecular design by combining deep autoencoder recurrent neural networks with generative topographic mapping. *Journal of chemical information and modeling*, 59(3):1182–1196, 2019.

[15] Zhenpeng Yao, Benjamin Sanchez-Lengeling, N Scott Bobbitt, Benjamin J Bucior, Sai Govind Hari Kumar, Sean P Collins, Thomas Burns, Tom K Woo, Omar Farha, Randall Q Snurr, et al. Inverse design of nanoporous crystalline reticular materials with deep generative models. 2020.

[16] Abdulelah S Alshehri, Rafiqul Gani, and Fengqi You. Deep learning and knowledge-based methods for computer aided molecular design–toward a unified approach: State-of-the-art and future directions. *arXiv preprint arXiv:2005.08968*, 2020.

[17] Jacques Boitreaud, Vincent Mallet, Carlos Oliver, and Jerome Waldispuhl. Optimol: Optimization of binding affinities in chemical space for drug discovery. *bioRxiv*, 2020.

[18] Austin Tripp, Erik Daxberger, and José Miguel Hernández-Lobato. Sample-efficient optimization in the latent space of deep generative models via weighted retraining. *arXiv preprint arXiv:2006.09191*, 2020.

[19] Wengong Jin, Regina Barzilay, and Tommi Jaakkola. Hierarchical generation of molecular graphs using structural motifs. *arXiv preprint arXiv:2002.03230*, 2020.

[20] Wengong Jin, Regina Barzilay, and Tommi Jaakkola. Multi-objective molecule generation using interpretable substructures, 2020.

[21] Philippe Schwaller, Theophile Gaudin, David Lanyi, Costas Bekas, and Teodoro Laino. "Found in Translation": predicting outcomes of complex organic chemistry reactions using neural sequence-to-sequence models. *Chemical Science*, 9(28):6091–6098, 2018.

[22] Wengong Jin, Connor Coley, Regina Barzilay, and Tommi Jaakkola. Predicting organic reaction outcomes with Weisfeiler-Lehman network. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems 30*, pages 2607–2616. 2017.

[23] Bowen Liu, Bharath Ramsundar, Prasad Kawthekar, Jade Shi, Joseph Gomes, Quang Luu Nguyen, Stephen Ho, Jack Sloane, Paul Wender, and Vijay Pande. Retrosynthetic reaction prediction using neural sequence-to-sequence models. *ACS central science*, 3(10): 1103–1113, 2017.

[24] Ryan-Rhys Griffiths, Philippe Schwaller, and Alpha A. Lee. Dataset bias in the natural sciences: A case study in chemical reaction prediction and synthesis design. *ChemRxiv*, 2018.

[25] Giorgio Pesciullesi, Philippe Schwaller, Teodoro Laino, and Jean-Louis Reymond. Carbohydrate transformer: Predicting regio-and stereoselective reactions using transfer learning. 2020.

[26] Yao Zhang et al. Bayesian semi-supervised learning for uncertainty-calibrated prediction of molecular properties and active learning. *Chemical Science*, 10(35):8154–8163, 2019.

[27] Seongok Ryu, Yongchan Kwon, and Woo Youn Kim. A Bayesian graph convolutional network for reliable prediction of molecular properties with uncertainty quantification. *Chemical Science*, 10(36):8438–8446, 2019.

[28] Zhenqin Wu, Bharath Ramsundar, Evan N. Feinberg, Joseph Gomes, Caleb Geniesse, Aneesh S. Pappu, Karl Leswing, and Vijay Pande. MoleculeNet: A benchmark for molecular machine learning. *Chemical Science*, 9(2):513–530, 2018. doi: 10.1039/C7SC02664A.

[29] Soojung Yang, Kyung Hoon Lee, and Seongok Ryu. A comprehensive study on the prediction reliability of graph neural networks for virtual screening. *arXiv preprint arXiv:2003.07611*, 2020.

[30] Wengong Jin, Regina Barzilay, and Tommi Jaakkola. Adaptive invariance for molecule property prediction. *arXiv preprint arXiv:2005.03004*, 2020.

[31] Katharina Hull, Johannes Morstein, and Dirk Trauner. In vivo photopharmacology. *Chemical reviews*, 118(21):10710–10747, 2018.

[32] Willem A Velema, Wiktor Szymanski, and Ben L Feringa. Photopharmacology: beyond proof of principle. *Journal of the American Chemical Society*, 136(6):2178–2191, 2014.

[33] Michael Wegener, Mickel J Hansen, Arnold JM Driessen, Wiktor Szymanski, and Ben L Feringa. Photocontrol of antibacterial activity: shifting from uv to red light activation. *Journal of the American Chemical Society*, 139(49):17979–17986, 2017.

[34] Matthew J Fuchter. On the promise of photopharmacology using photoswitches: a medicinal chemist's perspective. *Journal of Medicinal Chemistry*, 2020.

[35] Zhihang Wang, Raul Losantos, Diego Sampedro, Masa-aki Morikawa, Karl Börjesson, Nobuo Kimizuka, and Kasper Moth-Poulsen. Demonstration of an azobenzene derivative based solar thermal energy storage system. *Journal of Materials Chemistry A*, 7(25):15042–15047, 2019.

[36] Liqi Dong, Yiyu Feng, Ling Wang, and Wei Feng. Azobenzene-based solar thermal fuels: design, properties, and applications. *Chemical Society Reviews*, 47(19):7339–7368, 2018.

[37] Mihael A Gerkman, Rosina SL Gibson, Joaquín Calbo, Yuran Shi, Matthew J Fuchter, and Grace GD Han. Arylazopyrazoles for long-term thermal energy storage and optically-triggered heat release below 0 degrees c. *Journal of the American Chemical Society*, 2020.

[38] David L. Mobley and J. Peter Guthrie. FreeSolv: A database of experimental and calculated hydration free energies, with input files. *Journal of Computer-Aided Molecular Design*, 28(7): 711–720, July 2014. ISSN 0920-654X, 1573-4951. doi: 10.1007/s10822-014-9747-x.

[39] John S Delaney. Esol: estimating aqueous solubility directly from molecular structure. *Journal of chemical information and computer sciences*, 44(3):1000–1005, 2004.

[40] L. C. Blum and J.-L. Reymond. 970 million druglike small molecules for virtual screening in the chemical universe database GDB-13. *J. Am. Chem. Soc.*, 131:8732, 2009.

[41] M. Rupp, A. Tkatchenko, K.-R. Müller, and O. A. von Lilienfeld. Fast and accurate modeling of molecular atomization energies with machine learning. *Physical Review Letters*, 108: 058301, 2012.

[42] Grégoire Montavon, Matthias Rupp, Vivekanand Gobre, Alvaro Vazquez-Mayagoitia, Katja Hansen, Alexandre Tkatchenko, Klaus-Robert Müller, and O Anatole von Lilienfeld. Machine learning of molecular electronic properties in chemical compound space. *New Journal of Physics*, 15(9):095003, 2013. URL http://stacks.iop.org/1367-2630/15/i=9/a=095003.

[43] Lars Ruddigkeit, Ruud Van Deursen, Lorenz C Blum, and Jean-Louis Reymond. Enumeration of 166 billion organic small molecules in the chemical universe database gdb-17. *Journal of chemical information and modeling*, 52(11):2864–2875, 2012.

[44] Raghunathan Ramakrishnan, Pavlo O Dral, Matthias Rupp, and O Anatole von Lilienfeld. Big data meets quantum chemistry approximations: The δ-machine learning approach. *Journal of chemical theory and computation*, 11(5):2087–2096, 2015.

[45] Raghunathan Ramakrishnan, Pavlo O Dral, Matthias Rupp, and O Anatole von Lilienfeld. Quantum chemistry structures and properties of 134 kilo molecules. *Scientific Data*, 1, 2014.

[46] Asim Kumar Debnath, Rosa L Lopez de Compadre, Gargi Debnath, Alan J Shusterman, and Corwin Hansch. Structure-activity relationship of mutagenic aromatic and heteroaromatic nitro compounds. correlation with molecular orbital energies and hydrophobicity. *Journal of medicinal chemistry*, 34(2):786–797, 1991.

[47] Kristian Kersting, Nils M. Kriege, Christopher Morris, Petra Mutzel, and Marion Neumann. Benchmark data sets for graph kernels, 2016.

[48] Ann M Richard, Richard S Judson, Keith A Houck, Christopher M Grulke, Patra Volarath, Inthirany Thillainadarajah, Chihae Yang, James Rathman, Matthew T Martin, John F Wambaugh, et al. Toxcast chemical landscape: paving the road to 21st century toxicology. *Chemical research in toxicology*, 29(8):1225–1251, 2016.

[49] Cheng-Wei Ju, Hanzhi Bai, Rizhang Liu, and Bo Li. Can Machine Learning Be More Accurate Than TD-DFT? Prediction of Emission Wavelengths and Quantum Yields of Organic Fluorescent Materials. 4 2020. doi: 10.26434/chemrxiv.12111060.v1.

[50] Kirill Degtyarenko, Paula De Matos, Marcus Ennis, Janna Hastings, Martin Zbinden, Alan McNaught, Rafael Alcántara, Michael Darsow, Mickaël Guedj, and Michael Ashburner. Chebi: a database and ontology for chemical entities of biological interest. *Nucleic acids research*, 36 (suppl_1):D344–D350, 2007.

[51] Sunghwan Kim, Paul A Thiessen, Evan E Bolton, Jie Chen, Gang Fu, Asta Gindulyte, Lianyi Han, Jane He, Siqian He, Benjamin A Shoemaker, et al. Pubchem substance and compound databases. *Nucleic acids research*, 44(D1):D1202–D1213, 2016.

[52] Yanli Wang, Stephen H Bryant, Tiejun Cheng, Jiyao Wang, Asta Gindulyte, Benjamin A Shoemaker, Paul A Thiessen, Siqian He, and Jian Zhang. Pubchem bioassay: 2017 update. *Nucleic acids research*, 45(D1):D955–D963, 2017.

[53] Edward J Beard, Ganesh Sivaraman, Álvaro Vázquez-Mayagoitia, Venkatram Vishwanath, and Jacqueline M Cole. Comparative dataset of experimental and computational attributes of uv/vis absorption spectra. *Scientific Data*, 6(1):1–11, 2019.

[54] Marc Brockschmidt. {GNN}-fi{lm}: Graph neural networks with feature-wise linear modulation, 2020. URL https://openreview.net/forum?id=HJe4Cp4KwH.

[55] Leland McInnes, John Healy, Nathaniel Saul, and Lukas Großberger. Umap: Uniform manifold approximation and projection. *Journal of Open Source Software*, 3(29):861, 2018.

[56] Hyunjik Kim, Andriy Mnih, Jonathan Schwarz, Marta Garnelo, Ali Eslami, Dan Rosenbaum, Oriol Vinyals, and Yee Whye Teh. Attentive neural processes. In *International Conference on Learning Representations*, 2019. URL https://arxiv.org/abs/1901.05761.

[57] Thomas N. Kipf and Max Welling. Semi-supervised classification with graph convolutional networks. In *International Conference on Learning Representations (ICLR)*, 2017. URL https://arxiv.org/abs/1609.02907.

[58] Petar Veličković, Guillem Cucurull, Arantxa Casanova, Adriana Romero, Pietro Liò, and Yoshua Bengio. Graph Attention Networks. *In International Conference on Learning Representations*, 2018. URL https://arxiv.org/abs/1710.10903.

[59] Kevin Yang, Kyle Swanson, Wengong Jin, Connor Coley, Philipp Eiden, Hua Gao, Angel Guzman-Perez, Timothy Hopper, Brian Kelley, Miriam Mathea, et al. Analyzing learned molecular representations for property prediction. *Journal of chemical information and modeling*, 59(8):3370–3388, 2019.

[60] David Rogers and Mathew Hahn. Extended-connectivity fingerprints. *Journal of chemical information and modeling*, 50(5):742–754, 2010.

[61] Greg Landrum et al. Rdkit: Open-source cheminformatics. 2006. URL http://www.rdkit.org.

[62] Albert P Bartók, Risi Kondor, and Gábor Csányi. On representing chemical environments. *Physical Review B*, 87(18):184115, 2013.

[63] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

[64] Alexander G De G. Matthews, Mark Van Der Wilk, Tom Nickson, Keisuke Fujii, Alexis Boukouvalas, Pablo León-Villagrá, Zoubin Ghahramani, and James Hensman. Gpflow: A gaussian process library using tensorflow. *The Journal of Machine Learning Research*, 18(1): 1299–1304, 2017.

[65] Liva Ralaivola, Sanjay J Swamidass, Hiroto Saigo, and Pierre Baldi. Graph kernels for chemical informatics. *Neural networks*, 18(8):1093–1110, 2005.

[66] D. Kingma and J. Ba. Adam: A method for stochastic optimization. *arXiv preprint*. URL https://arxiv.org/abs/1412.6980.

[67] Minjie Wang, Lingfan Yu, Da Zheng, Quan Gan, Yu Gai, Zihao Ye, Mufei Li, Jinjing Zhou, Qi Huang, Chao Ma, Ziyue Huang, Qipeng Guo, Hao Zhang, Haibin Lin, Junbo Zhao, Jinyang Li, Alexander J Smola, and Zheng Zhang. Deep graph library: Towards efficient and scalable deep learning on graphs. *ICLR Workshop on Representation Learning on Graphs and Manifolds*, 2019. URL https://arxiv.org/abs/1909.01315.

[68] Sergey Ioffe and Christian Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *Proceedings of the 32nd International Conference on International Conference on Machine Learning - Volume 37*, pages 448–456. JMLR.org, 2015.

[69] Lauri Himanen, Marc OJ Jäger, Eiaki V Morooka, Filippo Federici Canova, Yashasvi S Ranawat, David Z Gao, Patrick Rinke, and Adam S Foster. Dscribe: Library of descriptors for machine learning in materials science. *Computer Physics Communications*, 247:106949, 2020.

[70] Justin Gilmer, Samuel S Schoenholz, Patrick F Riley, Oriol Vinyals, and George E Dahl. Neural message passing for quantum chemistry. In *Proceedings of the 34th International Conference on Machine Learning - Volume 70*, pages 1263–1272. JMLR.org, 2017.

[71] José Miguel Hernández-Lobato, Yingzhen Li, Mark Rowland, Daniel Hernández-Lobato, Thang D Bui, and Richard E Turner. Black-box α-divergence minimization. In *Proceedings of the 33rd International Conference on International Conference on Machine Learning-Volume 48*, pages 1511–1520, 2016.

[72] Guillaume Lambard and Ekaterina Gracheva. Smiles-x: autonomous molecular compounds characterization for small datasets without descriptors. *Machine Learning: Science and Technology*, 1(2):025004, 2020.

[73] Takeshi Yanai, David P Tew, and Nicholas C Handy. A new hybrid exchange–correlation functional using the coulomb-attenuating method (cam-b3lyp). *Chemical Physics Letters*, 393 (1-3):51–57, 2004.

[74] John P Perdew, Matthias Ernzerhof, and Kieron Burke. Rationale for mixing exact exchange with density functional approximations. *The Journal of chemical physics*, 105(22):9982–9985, 1996.

[75] Carlo Adamo and Vincenzo Barone. Toward reliable density functional methods without adjustable parameters: The pbe0 model. *The Journal of chemical physics*, 110(13):6158–6170, 1999.

[76] Denis Jacquemin, Julien Preat, Eric A Perpète, Daniel P Vercauteren, Jean-Marie André, Ilaria Ciofini, and Carlo Adamo. Absorption spectra of azobenzenes simulated with time-dependent density functional theory. *International Journal of Quantum Chemistry and references*, 111 (15):4224–4240, 2011.

[77] 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.

[78] Joaquin Calbo, Aditya R Thawani, Rosina SL Gibson, Andrew JP White, and Matthew J Fuchter. A combinatorial approach to improving the performance of azoarene photoswitches. *Beilstein Journal of Organic Chemistry*, 15(1):2753–2764, 2019.

[79] Joaquín Calbo, Claire E Weston, Andrew JP White, Henry S Rzepa, Julia Contreras-García, and Matthew J Fuchter. Tuning azoheteroarene photoswitch performance through heteroaryl design. *Journal of the American Chemical Society*, 139(3):1261–1274, 2017.

[80] Claire E Weston, Robert D Richardson, Peter R Haycock, Andrew JP White, and Matthew J Fuchter. Arylazopyrazoles: azoheteroarene photoswitches offering quantitative isomerization and long thermal half-lives. *Journal of the American Chemical Society*, 136(34):11878–11881, 2014.

[81] Ron Siewertsen, Hendrikje Neumann, Bengt Buchheim-Stehn, Rainer Herges, Christian Nather, Falk Renth, and Friedrich Temps. Highly efficient reversible z- e photoisomerization of a bridged azobenzene with visible light through resolved s1 (npi*) absorption bands. *Journal of the American Chemical Society*, 131(43):15594–15595, 2009.

[82] Karin Rustler, Philipp Nitschke, Sophie Zahnbrecher, Julia Zach, Stefano Crespi, and Burkhard Konig. Photochromic evaluation of 3 (5)-arylazo-1 h-pyrazoles. *The Journal of Organic Chemistry*, 85(6):4079–4088, 2020.

[83] Christopher Knie, Manuel Utecht, Fangli Zhao, Hannes Kulla, Sergey Kovalenko, Albert M Brouwer, Peter Saalfrank, Stefan Hecht, and David Bléger. ortho-Fluoroazobenzenes: visible light switches with very long-lived Z isomers. *Chemistry–A European Journal*, 20(50): 16492–16501, 2014.

[84] Hanno Sell, Christian Näther, and Rainer Herges. Amino-substituted diazocines as pincer-type photochromic switches. *Beilstein journal of organic chemistry*, 9(1):1–7, 2013.

[85] Steffen Thies, Hanno Sell, Claudia Bornholdt, Christian Schütt, Felix Köhler, Felix Tuczek, and Rainer Herges. Light-driven coordination-induced spin-state switching: Rational design of photodissociable ligands. *Chemistry–A European Journal*, 18(51):16358–16368, 2012.

[86] Sudha Devi, Mayank Saraswat, Surbhi Grewal, and Sugumar Venkataramani. Evaluation of substituent effect in z-isomer stability of arylazo-1 h-3, 5-dimethylpyrazoles: Interplay of steric, electronic effects and hydrogen bonding. *The Journal of organic chemistry*, 83(8): 4307–4322, 2018.

[87] Pravesh Kumar, Anjali Srivastava, Chitranjan Sah, Sudha Devi, and Sugumar Venkataramani. Arylazo-3, 5-dimethylisoxazoles: Azoheteroarene photoswitches exhibiting high z-isomer stability, solid-state photochromism, and reversible light-induced phase transition. *Chemistry–A European Journal*, 25(51):11924–11932, 2019.

[88] Chavdar Slavov, Chong Yang, Andreas H Heindl, Hermann A Wegner, Andreas Dreuw, and Josef Wachtveitl. Thiophenylazobenzene: An alternative photoisomerization controlled by lone-pair pi interaction. *Angewandte Chemie*, 132(1):388–395, 2020.

[89] Denis Jacquemin, Julien Preat, Eric A Perpète, Daniel P Vercauteren, Jean-Marie André, Ilaria Ciofini, and Carlo Adamo. Absorption spectra of azobenzenes simulated with time-dependent density functional theory. *International Journal of Quantum Chemistry*, 111(15):4224–4240, 2011.

[90] Heinz Mustroph and Frank Gussmann. Studies on uv/vis absorption spectra of azo dyes. 24. the different effect of a 2-methoxy and a 3-methoxy group in 4-nn-diethylaminoazobenzenes on colour. *Journal für Praktische Chemie*, 332(1):93–97, 1990.

[91] Heinz Mustroph. Studies on the uv-vis absorption spectra of azo dyes: Part 25. analysis of the fine structure of the pi1-pi1* band of 4-donor-sub. *Dyes and pigments*, 15(2):129–137, 1991.

[92] Heinz Mustroph. Studies on uv/vis absorption spectra of azo dyes.: Part 26. electronic absorption spectra of 4, 4'-diaminoazobenzenes. *Dyes and pigments*, 16(3):223–230, 1991.

[93] I Bridgeman and AT Peters. Synthesis and electronic spectra of some 4–aminoazobenzenes. *Journal of the Society of Dyers and Colourists*, 86(12):519–524, 1970.

[94] Zeynel Seferoğlu, Nermin Ertan, Tuncer Hökelek, and Ertan Şahin. The synthesis, spectroscopic properties and crystal structure of novel, bis-hetarylazo disperse dyes. *Dyes and Pigments*, 77(3):614–625, 2008.

[95] Haluk Dinçalp, Sinem Yavuz, Özgül Haklı, Ceylan Zafer, Cihan Özsoy, İnci Durucasu, and Sıddık İçli. Optical and photovoltaic properties of salicylaldimine-based azo ligands. *Journal of photochemistry and Photobiology A: Chemistry*, 210(1):8–16, 2010.

[96] Aaron DW Kennedy, Isolde Sandler, Joakim Andréasson, Junming Ho, and Jonathon E Beves. Visible-light photoswitching by azobenzazoles. *Chemistry–A European Journal*, 26(5): 1103–1110, 2020.

[97] H Faustino, CR Brannigan, LV Reis, PF Santos, and P Almeida. Novel azobenzothiazole dyes from 2-nitrosobenzothiazoles. *Dyes and Pigments*, 83(1):88–94, 2009.

[98] Aytül Saylam, Zeynel Seferoğlu, and Nermin Ertan. Azo-8-hydroxyquinoline dyes: the synthesis, characterizations and determination of tautomeric properties of some new phenyl- and heteroarylazo-8-hydroxyquinolines. *Journal of Molecular Liquids*, 195:267–276, 2014.

[99] Ming Shien Yen and Jing Wang. Synthesis and absorption spectra of hetarylazo dyes derived from coupler 4-aryl-3-cyano-2-aminothiophenes. *Dyes and Pigments*, 61(3):243–250, 2004.

[100] Felix A Faber, Luke Hutchison, Bing Huang, Justin Gilmer, Samuel S Schoenholz, George E Dahl, Oriol Vinyals, Steven Kearnes, Patrick F Riley, and O Anatole Von Lilienfeld. Prediction errors of molecular machine learning models lower than hybrid DFT error. *Journal of chemical theory and computation*, 13(11):5255–5264, 2017.

[101] Anders S Christensen, Lars A Bratholm, Felix A Faber, and O Anatole von Lilienfeld. FCHL revisited: faster and more accurate quantum machine learning. *The Journal of Chemical Physics*, 152(4):044107, 2020.

[102] Theano Development Team. Theano: A Python framework for fast computation of mathematical expressions. *arXiv e-prints*, abs/1605.02688, May 2016. URL http://arxiv.org/abs/1605.02688.

[103] Veronika Brázdová and D. R. Bowler. *Atomistic computer simulations: a practical guide*. Wiley-VCH, Weinheim, 2013. ISBN 978-3-527-41069-9. OCLC: ocn835961914.

[104] Axel D Becke. Perspective: Fifty years of density-functional theory in chemical physics. *The Journal of chemical physics*, 140(18):18A301, 2014.

[105] Andrew R Leach and Andrew R Leach. *Molecular modelling: principles and applications*. Pearson education, 2001.

[106] Philip J Hasnip, Keith Refson, Matt IJ Probert, Jonathan R Yates, Stewart J Clark, and Chris J Pickard. Density functional theory in the solid state. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 372(2011):20130270, 2014.

[107] P Hoenberg and W Kohn. Inhomogeneous electron gas. *Phys. Rev*, 136:B864–B871, 1964.

[108] J Coleman Howard, Jordan D Enyard, and Gregory S Tschumper. Assessing the accuracy of some popular dft methods for computing harmonic vibrational frequencies of water clusters. *The Journal of chemical physics*, 143(21):214103, 2015.

[109] Habbo H Heinze, Andreas Görling, and Notker Rösch. An efficient method for calculating molecular excitation energies by time-dependent density-functional theory. *The Journal of Chemical Physics*, 113(6):2088–2099, 2000.

[110] Robert van Leeuwen. Causality and symmetry in time-dependent density-functional theory. *Physical review letters*, 80(6):1280, 1998.

[111] Carsten A Ullrich. *Time-dependent density-functional theory: concepts and applications*. OUP Oxford, 2011.

[112] Mark E Casida and Miquel Huix-Rotllant. Progress in time-dependent density-functional theory. *Annual review of physical chemistry*, 63:287–323, 2012.

[113] Kieron Burke, Jan Werschnik, and EKU Gross. Time-dependent density functional theory: Past, present, and future. *The Journal of chemical physics*, 123(6):062206, 2005.

A Details of Dataset Construction

A.1 Tabulated Properties

The dataset includes molecular properties for 405 photoswitch molecules in total. All molecular structures are denoted according to the simplified molecular input line entry system (SMILES) [77]. We collate the following molecular properties, where available:

- **Rate of Thermal Isomerization** (units = s^{-1}): This is a measure of the thermal stability of the least stable isomer (Z isomer for non-cyclic azophotoswitches and E isomer for cyclic azophotoswitches). Measurements are carried out in solution with the compounds dissolved in the stated solvents.
- **Photostationary State** (units = % of stated isomer): Upon continuous irradiation of an azophotoswitch a steady state distribution of the E and Z isomers is achieved. Measurements are carried out in solution with the compounds dissolved in the 'irradiation solvents'.
- **Experimental Transition Wavelengths** (units = nanometers): The wavelength at which the $\pi - \pi^*/n - \pi^*$ electronic transition has a maxima for the stated isomer. Measurements are carried out in solution with the compounds dissolved in the 'irradiation solvents'.
- **DFT-Computed Transition Wavelengths** (units = nanometers): DFT-computed wavelengths at which the $\pi - \pi^*/n - \pi^*$ electronic transition has a maxima for the stated isomer.
- **Extinction coefficient**: (units = $\mathrm{M}^{-1}\mathrm{cm}^{-1}$) A measure of how strongly a molecular species in solution absorbs light.
- **Wiberg Index**: A measure of the bond order of the N=N bond in an azophotoswitch. Bond order is a measure of the 'strength' of said chemical bond. This value is computed theoretically.
- **Irradiation wavelength**: The specific wavelength of light used to irradiate samples from E-Z or Z-E such that a photo stationary state is obtained. Measurements are carried out in solution with the compounds dissolved in the 'irradiation solvents'.
- **Irradiation Solvents**: The solvent used to obtain the aforementioned photophysical values.

A.2 Sources of Experimental Data

Properties in subsection A.1 were collated from a wide range of photoswitch literature. An emphasis was placed on collating compounds with a wide range of functional groups attached to the core photoswitch scaffold. In addition, this dataset is unique in that it is composed of the latest generations of azoheteroarenes and cyclic azobenzenes which possess far superior photoswitch properties to analogous, unmodified azobenezenes. See Figure 4 for an overview of these novel azophotoswitches with their properties summarised. [78, 79, 80, 81, 82, 83, 84, 85, 86, 87, 88, 89, 90, 91, 92, 93, 94, 95, 96, 97, 98, 99].

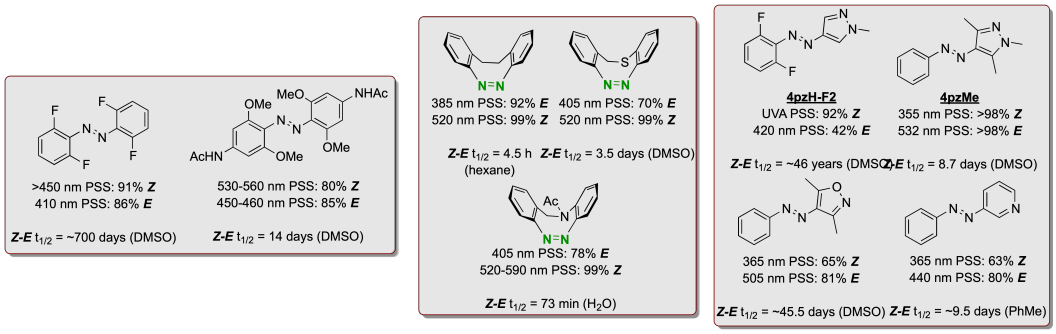


Figure 4: A data summary for the latest generation of azophotoswitches contained in this dataset. **PSS** = photostationary state, ***Z*-*E*** $t_{1/2}$ = *Z* isomer thermal half-life.

B Dataset Visualizations

The choice of molecular representation is known to be a key factor in the performance of machine learning algorithms on molecules [100, 28, 101]. Commonly-used representations such as fingerprint and fragment-based descriptors are high-dimensional and as such, it can be challenging to interpret the inductive bias introduced by the representation. In order to visualize the high-dimensional representation space of the Photoswitch Dataset we project the data matrix to two dimensions using the UMAP algorithm [55]. We compare the manifolds located under the Morgan fingerprint representation and a fragment-based representation computed using RDKit [61]. We generate 512-bit Morgan fingerprints with a bond radius of 2, setting the nearest neighbours parameter in the UMAP algorithm to a value of 50. The resulting visualization was produced using the ASAP package (available at https://github.com/BingqingCheng/ASAP) and is shown in Figure 5.

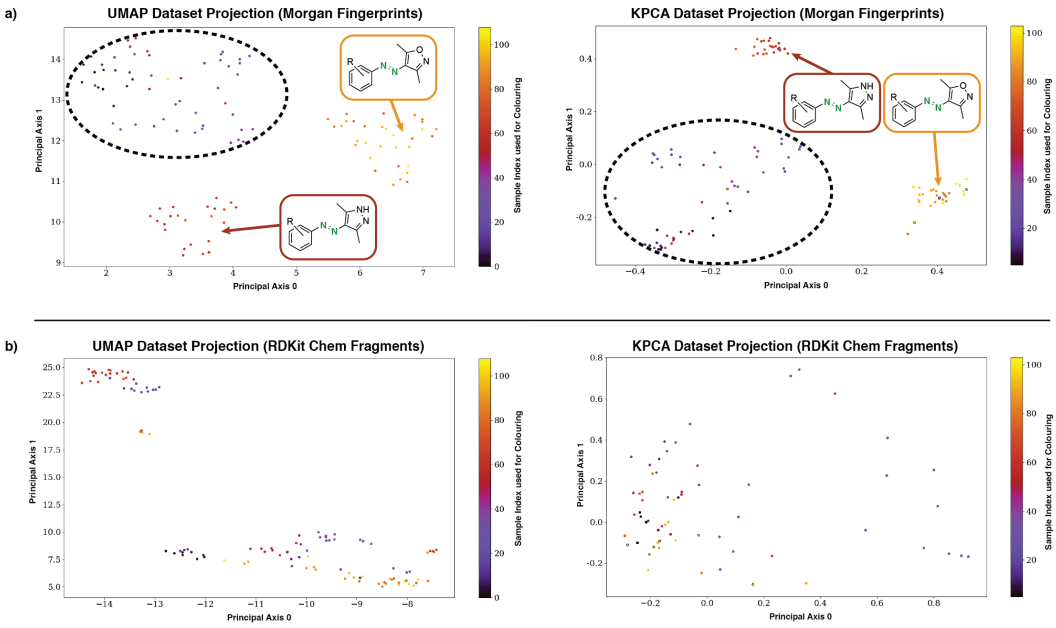


Figure 5: a) UMAP and k-PCA projections of the dataset, using Morgan Fingerprints, correctly identify clusters of chemically similar molecules. The regions demarcated by dashed black lines are composed of miscellaneous azoheteroarenes; no grouping was noted here due to the limited (≤ 10) examples per class. b) Similar projections using RDKit Fragment descriptors fails to identify any such clusters.

The structure of the manifold located under the Morgan fingerprint representation identifies meaningful subgroups of azophotoswitches when compared to the fragment-based representation. In order to demonstrate that the finding is due to the representation and not the dimensionality reduction algorithm we include the manifolds identified by k-PCA using a cosine kernel. Both algorithms identify the same manifold structure in the Morgan fingerprint representation.

C Further Experiments

C.1 Property Prediction

In this section we present, in Table 6 results with additional models on the property prediction benchmark from the main paper. The black-box alpha divergence minimization Bayesian neural network is implemented in the Theano library [102] and is based on the implementation of [71]. the network has 2 hidden layers of size 25 with ReLU activations. The alpha parameter is set to 0.5, the prior variance for the variational distribution q is set to 1 and 100 samples are taken to approximate the expectation over the variational distribution. For all tasks the network is trained using 8 iterations

of the ADAM optimizer [66] with a batch size of 32 and a learning rate of 0.05. The MPNN is trained for 100 epochs in the case of the E isomer $\pi - \pi^*$ task and 200 epochs in the case of the other tasks with a learning rate of 0.001 and a batch size of 32. The model architecture was taken to be the library default with the same node and edge features used for the GCN and GAT models in the main paper. The SMILES-X implementation remains the same as that of the paper [72] save for the difference that the network is trained for 40 epochs without Bayesian optimization over model architectures. In the case of SMILES-X 3 random train/test splits are used instead of 20 for the Z isomer tasks whereas 2 splits are used for the E isomer $n-\pi^*$ task. For the E isomer $\pi - \pi^*$ prediction task results are missing due to insufficient RAM on the machine used to run the experiments. The results presented here are withheld from the main paper because it is unclear as to whether the optimal model architecture has been identified in each case.

Table 6: Test Set Performance in Predicting the Transition Wavelengths of the E and Z isomers.

	E isomer $\pi - \pi^*$ (nm)	*E* isomer $n-\pi^*$ (nm)	Z isomer $\pi - \pi^*$ (nm)	*Z* isomer $n-\pi^*$ (nm)
RMSE				
BNN + Morgan	27.0 ± 0.9	12.9 ± 0.6	13.9 ± 0.6	12.7 ± 0.4
BNN + Fragments	31.2 ± 1.1	14.8 ± 0.8	16.9 ± 0.8	12.7 ± 0.4
BNN + Fragprints	26.7 ± 0.8	13.1 ± 0.5	14.9 ± 0.5	13.0 ± 0.6
MPNN	24.8 ± 0.8	12.5 ± 0.6	16.7 ± 0.8	12.8 ± 0.7
SMILES-X		25.1 ± 4.2	17.8 ± 0.6	14.8 ± 0.9
MAE				
BNN + Morgan	19.0 ± 0.6	9.9 ± 0.4	10.2 ± 0.5	8.6 ± 0.3
BNN + Fragments	22.4 ± 0.8	10.6 ± 0.4	12.9 ± 0.6	8.6 ± 0.3
BNN + Fragprints	19.1 ± 0.6	10.1 ± 0.5	10.8 ± 0.4	9.3 ± 0.5
MPNN	15.4 ± 0.8	8.6 ± 0.3	11.6 ± 0.6	8.4 ± 0.4
SMILES-X		20.6 ± 3.1	11.6 ± 1.0	11.2 ± 1.0
$\boldsymbol{R^2}$				
BNN + Morgan	0.83 ± 0.01	0.69 ± 0.02	0.23 ± 0.08	0.18 ± 0.05
BNN + Fragments	0.77 ± 0.01	0.58 ± 0.04	-0.15 ± 0.14	0.18 ± 0.05
BNN + Fragprints	0.83 ± 0.01	0.68 ± 0.02	0.14 ± 0.06	0.11 ± 0.08
MPNN	0.83 ± 0.01	0.63 ± 0.06	-0.70 ± 0.34	-0.68 ± 0.27
SMILES-X		-0.44 ± 0.30	-0.08 ± 0.06	-0.09 ± 0.04

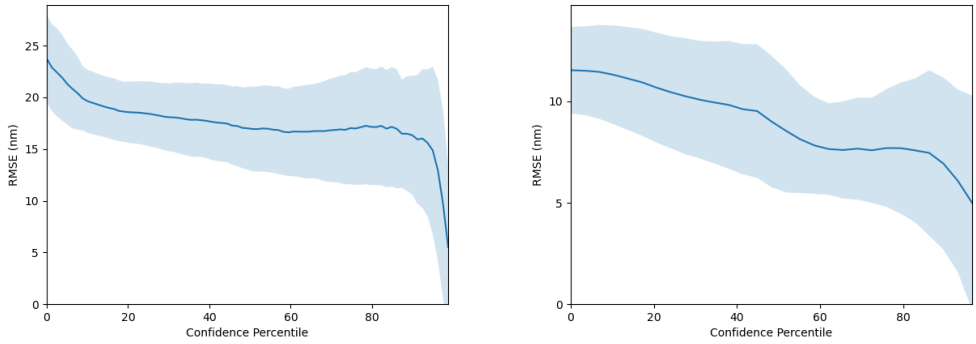
C.2 Confidence-Error Curves

An advantage of Bayesian models for the real-world prediction task is the ability to produce calibrated uncertainty estimates. If correlated with prediction error, a model's uncertainty may act as an additional decision-making criterion for the selection of candidates for lab synthesis. In order to investigate the benefits afforded by uncertainty estimates, we produce confidence-error curves using the GP-Tanimoto model in conjunction with the fingerprints representation. The confidence-error curves for the RMSE and MAE metrics are shown in Figure 6 and Figure 7 respectively. The x-axis, confidence percentile, may be obtained simply by ranking each model prediction of the test set in terms of the predictive variance at the location of that test input. As an example, molecules that lie in the 80th confidence percentile will be the 20% of test set molecules with the lowest model uncertainty. We then measure the prediction error at each confidence percentile across 200 random train/test splits to see whether the model's confidence is correlated with the prediction error. We observe that across all tasks the GP-Tanimoto model's uncertainty estimates are positively correlated with prediction error, offering a proof of concept that model uncertainty can be incorporated into the decision process for candidate selection.

D Background on Time-Dependent Density Functional Theory

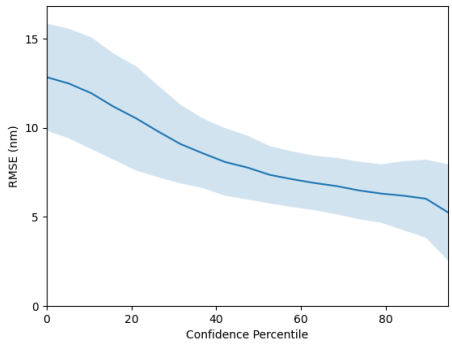
D.1 Density Functional Theory

Density Functional Theory (DFT) is a modelling method used to elucidate the electronic structure (typically the ground state) of many-body systems [103]. The theory has been used with great success across physics, chemistry, biology and materials science [104]. DFT is considered to be an *ab initio*, or first principles method because it relies directly upon the postulates of quantum mechanics and the only inputs to the calculations are physical constants [105]. A concrete example of an application of DFT towards an electronic structure investigation is in simulating a relaxation of atoms in a crystalline

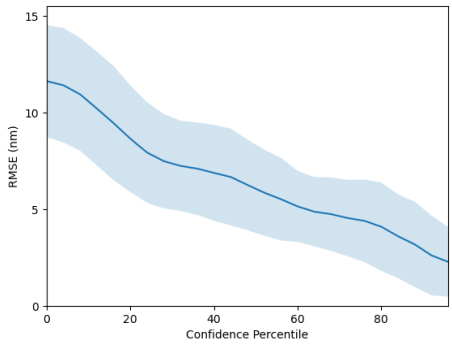


(a) E Isomer $\pi - \pi^*$

(b) E Isomer $n-\pi^*$



(c) Z Isomer $\pi - \pi^*$



(d) Z Isomer $n-\pi^*$

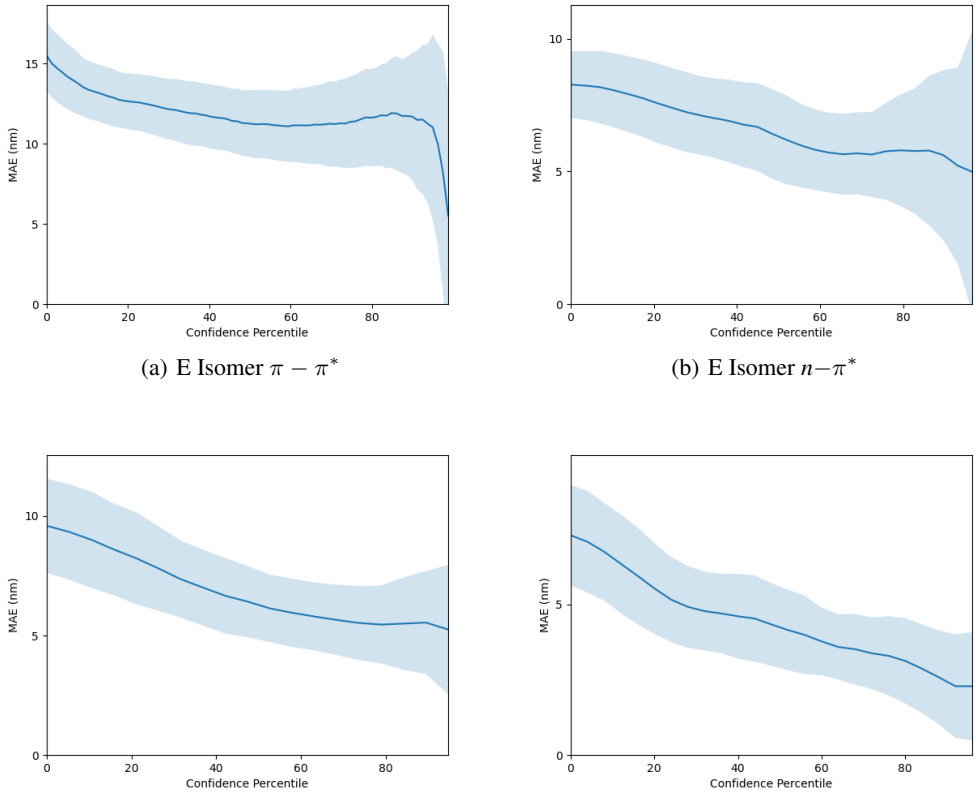
Figure 6: RMSE Confidence-Error Curves for Property Prediction using GP Regression.

solid to calculate the change in lattice parameters and the forces on each atom, with the introduction of defects or vacancies into the system[106].

Since its inception in 1964/5, Kohn-Sham DFT (KS-DFT) has been one of the most popular electronic structure methods to date [104]. KS-DFT relies on the Hohenberg-Kohn theorems [107] and the use of a trial electron density (an initial guess) with a self-consistency scheme. In practice, a computational loop takes a trial density, solves the Kohn-Sham equations, and obtains the single electron wavefunctions corresponding to the trial density; next, by taking these single electron wavefunctions and using a result of quantum mechanics, a calculated electron density can be computed. If this calculated density is consistent (within a set tolerance) of the trial density, then the theoretical ground state density has been found. If the two densities are not consistent, the calculated density is taken as the new trial density, and the loop is repeated until the tolerance is met. With exchange and correlation functionals, the accuracy of DFT calculations can be very high, but may also fluctuate significantly with the choice of functional, pseudopotential, basis sets and cutoff energy [108] which are not always straightforward to optimize. A machine learning corollary would be the performance of a specific model, on a given dataset, greatly depending on its hyperparameters, with out-of-the-box implementations rarely giving satisfactory results without a significant amount of tuning.

D.2 Time-Dependent Density Functional Theory

Time-dependent Density Functional Theory (TD-DFT) is based on a time-dependent cognate of the Hohenberg-Kohn theorems; the Runge-Gross (RG) theorem [109]. This theorem shows that a unique delineation exists between the time-dependent electron density and the time-dependent external potential. This allows for a simplification, permitting a computational time-dependent Kohn-Sham system to be substantiated [110] analogous to the computational system used in KS-DFT.



(a) E Isomer $\pi - \pi^*$

(b) E Isomer $n-\pi^*$

(c) Z Isomer $\pi - \pi^*$

(d) Z Isomer $n-\pi^*$

Figure 7: MAE Confidence-Error Curves for Property Prediction using GP Regression.

In conjunction with a linear response theory [111], TD-DFT has excelled with investigations into calculating electromagnetic spectra, i.e. absorption spectra, of medium and large molecules [112]. [113] It has become popular in these fields, due to its ease of use relative to other methods as well as its high accuracy. A relevant application of this methodology is to compute the $\pi - \pi^*/n - \pi^*$ electronic transitions wavelengths for conjugated molecular systems, such as the photoswitch molecules in the Photoswitch Dataset.