Combining Quantum Mechanical Calculations with Machine Learning and Genetic Algorithms for the Design of Better Materials

by

Omri D. Abarbanel

Bachelor of Arts, City University of New York - Hunter College, 2017

Submitted to the Graduate Faculty of

the Dietrich School of Arts and Sciences in partial fulfillment

of the requirements for the degree of

Doctor of Philosophy

University of Pittsburgh

2024

UNIVERSITY OF PITTSBURGH

DIETRICH SCHOOL OF ARTS AND SCIENCES

This dissertation was presented

by

Omri D. Abarbanel

It was defended on

August 6th, 2024

and approved by

Geoffrey Hutchison, Department of Chemistry

Peng Liu, Department of Chemistry

David Waldeck, Department of Chemistry

Olexander Isayev, Carnegie Mellon University, Department of Chemistry

Combining Quantum Mechanical Calculations with Machine Learning and Genetic Algorithms for the Design of Better Materials

Omri D. Abarbanel, PhD

University of Pittsburgh, 2024

In the past, the discovery process of new materials was done mainly through trial and error, which was time-consuming and expensive. However, computational simulations and models can quickly filter through a large number of potential candidates and narrow down the search space efficiently and cost-effectively. For example, this process can help identify new electronic materials that use less energy or have novel properties that open the door for new applications and find life-saving drugs for hard-to-cure or rare diseases.

In this work, we present how the combination of several of these computational techniques, namely quantum mechanical (QM) calculations with machine learning (ML) and Genetic Algorithms (GA), can help accelerate the discovery of new materials. We have used GFN2-xTB throughout this work because it has a good balance of accuracy and speed and shows how it can be used as a surrogate for the more costly density functional theory (DFT) and how it can be used to generate molecular features for ML applications.

Three different molecular properties were selected to show how the combination of QM with ML and GAs is greater than the sum of its parts. First, we used GFN2-xTB to calculate geometrical features for a random forest ML algorithm to identify new thiophene-based π-conjugated polymers with low reorganization energies, achieving a RMSE of 0.036 eV and a speed-up of ∼13× over DFT. Second, we used GFN2-xTB calculations in a GA to help identify novel π-conjugated polymers with stable triplet ground states, finding more than 1,400 potential candidates. Finally, we present QupKake, a graph-neural-networks based ML model that used GFN2-xTB calculated features to predict the micro-pK_a of drug-like molecules, achieving a 30% improvement over existing models.

Table of Contents

List of Tables

List of Figures

Preface

My Ph.D. journey at Pitt was convoluted and had many hardships and low points, but as I near the end point, I look back and feel pride in my work and the path that brought me here. I started as a synthetic chemist working on macrocycle synthesis in a different lab at Pitt, but due to physical and mental health reasons, after 1.5 years I decided this was not the best path. Fortunately, I had an extensive programming background, and the transition into computational chemistry was relatively smooth. I am very grateful that my advisor, Dr. Geoffrey Hutchison, agreed to take me in, mentor me, and guide me toward this thesis. Knowing Python and Bash did help at the beginning, but having to learn a lot of new computational chemistry concepts was challenging. But Geoff was always there to help, explain and inspire and I cannot thank him enough. I joined his group on March 2nd 2020 and just two weeks later we went into quarentine due to the COVID-19 pandemic. This was a tough time for everyone and I am thankful that Geoff cared a lot about the well-being of his students and helped us all do our research remotely. Thank you to my committee, Dr. Peng Liu, Dr. David Waldeck, and Dr. Olexander Isayev, for their time and support. I would also like to thank my colleagues and friends from the Hutchison group, especially Dr. Dakota Folmsbee, Dr. Danielle Elsey, and Dr. Brianna Greenstein, for their support and friendship throughout the last 4 years.

I would like to thank the other people in my life, especially my parents and family, who supported me from afar with every decision, tried to understand what my research is about, and helped me get to where I am today. They supported my scientific endeavors since I was a young child who dreamed of becoming a scientist. From taking me to after-school science classes, sending me to the Israeli Arts and Science Academy boarding school away from home, and supporting my move to a new country at just 22 years old, both mentally and financially. I am proud to say that thanks to them I finally fulfilled my childhood dream and that I can call myself a scientist.

Pittsburgh, August 2024 O. Abarbanel

1.0 Introduction

The pursuit of novel materials with superior properties is a critical endeavor in chemistry and materials science. This thesis presents an integrated approach combining quantum mechanical calculations, machine learning, and genetic algorithms to design and discover new materials with desirable electronic and molecular properties. The main focus is on π-conjugated polymers, which have widespread applications in organic electronics, such as solar cells, transistors, and light-emitting devices.[7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17] These polymers are particularly attractive due to their tunable electronic properties and flexibility. Additionally, the investigation includes micro-pK_a predictions for drug-like molecules, which are crucial for understanding the behavior of pharmaceutical compounds in biological systems. Accurate prediction of micro-pK_a values is essential for drug design, as it influences the drug's solubility, absorption, distribution, and excretion.[18, 19, 20, 18, 19]

Recent advances in computational chemistry and artificial intelligence have enabled the development of sophisticated models that can predict material properties with high accuracy.[21, 22, 23, 24, 25, 26, 27, 28, 29, 30, 31, 5, 6, 32, 33] These advancements allow for the exploration of vast chemical spaces more efficiently, facilitating the discovery of materials that were previously inaccessible through traditional experimental methods. This thesis aims to demonstrate the potential of these computational techniques in accelerating the design and optimization of π-conjugated polymers and drug-like molecules. The integration of these techniques not only speeds up the discovery process, but also reduces costs and resource consumption, making it an attractive approach for both academia and industry.

1.1 π-Conjugated Polymers

Conjugated polymers are a class of organic materials characterized by alternating single and double bonds along their backbone, allowing for extensive delocalization of π-electrons. This delocalization imparts unique electronic and optical properties to these materials, mak-

ing them highly attractive for a variety of applications in organic electronics. Their versatility and tunability stem from the ability to modify their chemical structure, which can lead to changes in properties such as bandgap, charge mobility, and photoluminescence. The ability to fine-tune these properties makes π-conjugated polymers ideal candidates for use in devices like organic photovoltaics (OPVs),[7, 8, 9] chemical sensors,[10, 11, 12, 13, 14] and organic field-effect transistors (OFETs).[34, 35]

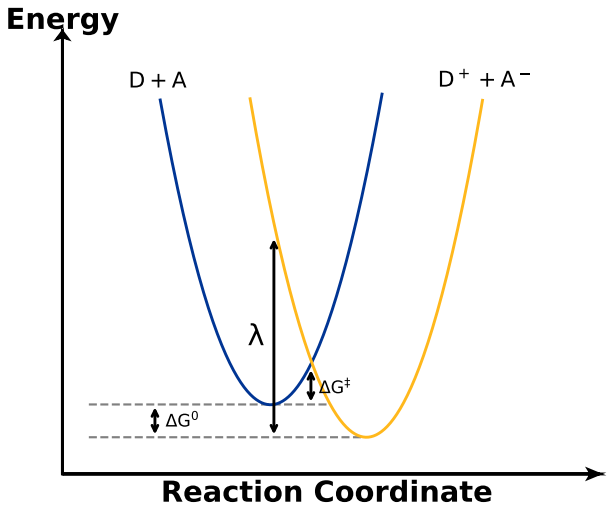


Figure 1.1: Reorganization energy (λ) is the energy required to facilitate a charge transfer reaction between a donor (D) and acceptor (A) molecules from their relaxed nuclear configuration (in blue) to the relaxed nuclear configuration of the products (in yellow.)

1.1.1 Reorganization Energy

Reorganization energy (λ) is an important parameter in the study of π-conjugated polymers, particularly for applications in organic electronics. It refers to the energy required to reorganize the molecular structure and its surroundings when a charge is added or removed (Figure 1.1). This property significantly influences charge transport properties, such as electron- and hole-mobility, which are essential for the performance of electronic devices like organic solar cells and transistors. Lower reorganization energy typically cor-

responds to higher charge mobility, leading to more efficient charge transport and better device performance.[29, 28, 36, 37, 38]

In π-conjugated polymers, the reorganization energy is influenced by the molecular structure, including factors such as conjugation length, planarity, and the presence and identity of substituents.[29, 28, 36, 37, 38] Computational techniques, including quantum mechanical methods and machine learning models, can predict reorganization energy with high accuracy. These predictions help to design polymers with optimal electronic properties for specific applications. For instance, polymers with low reorganization energy can be identified and synthesized for use in high-performance organic photovoltaics, enhancing their efficiency and stability.

1.1.2 Triplet Ground State

The ground state triplet is another important property of π-conjugated polymers, particularly relevant in the context of organic light-emitting diodes (OLEDs) and other optoelectronic and spintronic devices.[16, 17] The triplet state refers to a molecular electronic state with two unpaired electrons in two singly-occupied molecular orbitals (SOMOs), dubbed T_0, as opposed to a singlet ground state (S_0) with a pair of electrons with opposite spins occupying the same molecular orbital (HOMO).

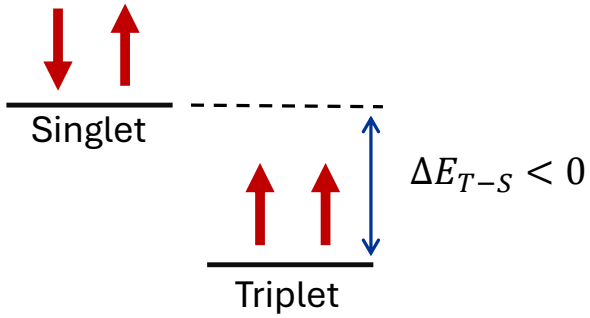


Figure 1.2: The triplet ground state is stable ($\Delta E_{T-S} < 0$) when the energy of the triplet species (E_T) is lower than the energy of the singlet species (E_S).

Understanding the conditions that leads to a triplet ground state in π-conjugated poly-

mers is crucial to designing materials with high triplet state stability. Quantum mechanical methods, such as DFT, can be used to calculate the energy levels and properties of triplet states. These calculations provide insights into the stability of the triplet states, which is defined as follows:

$$\Delta E_{T-S} = E_T - E_S \tag{1}$$

Where E_T is the energy of the triplet species, and E_S is the energy of the singlet species. When $\Delta E_{T-S} < 0$, the triplet ground state is more stable than the singlet ground state (Figure 1.2).

1.2 Micro-pK_a

The pK_a of a molecule is a fundamental property that significantly impacts its behavior in biological systems, including solubility, absorption, distribution, and excretion.[20, 18, 19] Accurate prediction of pK_a values is crucial for drug design and development, as it influences the drug's ionization state at different pH levels, affecting its pharmacokinetics and pharmacodynamics.[18, 19] Understanding the micro-pK_a of drug-like molecules allows researchers to design compounds with optimal properties for oral bioavailability and therapeutic efficacy. pK_a also affects the environmental impact of the materials, such as toxicity, reactivity, polymer solubility, and more.[39, 40, 41, 42, 43, 44]

Various quantum mechanical methods and machine learning models have previously been employed to predict micro-pK_a values with a root mean square error (RSME) of $\sim$1 pK_a unit.[45, 46, 47, 48, 49, 50] These computational techniques consider various factors, including the molecular structure, electronic environment, and solvation effects, to provide pK_a prediction and calculations. By integrating these predictions into the drug design process, researchers can systematically explore the chemical space of potential drug candidates and identify compounds with favorable pK_a values.

1.3 Inverse Design

Inverse design is a paradigm shift in material discovery, in which the desired properties of a material are specified first, and then the structure that achieves these properties is identified. This approach contrasts with the traditional trial-and-error method, in which materials are synthesized and characterized to determine their properties. Inverse design leverages computational techniques to predict the structures of materials that meet predefined criteria, significantly reducing the time and cost associated with material discovery. This method allows researchers to focus their efforts on the most promising candidates, thereby enhancing the efficiency and success rate of the discovery process. [51, 52]

The concept of inverse design is particularly powerful in the context of π-conjugated polymers and drug-like molecules. These materials exhibit a wide range of electronic and chemical properties that can be tuned by modifying their molecular structures. By using inverse design, researchers can target specific properties, such as HOMO-LUMO gap, reorganization energy, or micro-pK_a values, and design molecules that meet these targets. This approach allows for a more focused and efficient exploration of the chemical space, leading to the discovery of materials with optimal performance for specific applications.

One of the key challenges in inverse design is the accurate prediction of material properties from their molecular structures. This requires sophisticated computational models that can capture the complex relationships between structure and properties. Quantum mechanical methods, such as density functional theory (DFT), play a crucial role in this process by providing detailed insights into the electronic structure of materials. In addition, machine learning models is being used to predict molecular properties with high accuracy and speed, enabling rapid identification of potential candidates. This combination of computational techniques represents a powerful toolkit for modern material and computational scientists.

1.4 Quantum Mechanical Methods

QM methods are fundamental to the study of material properties at the molecular level. These methods, based on the principles of quantum mechanics, allow for the accurate calculation of electronic properties, such as energy levels, charge distribution, and molecular orbitals. DFT is one of the most widely used QM methods in material science because of its balance between accuracy and computational cost. DFT provides information on the electronic structure, which is crucial for predicting the behavior of materials under various conditions.

In the context of π-conjugated polymers and drug-like molecules, QM methods are essential to understand the relationship between molecular structure and properties. For example, the energy difference between the highest occupied molecular orbital and the lowest unoccupied molecular orbital (HOMO-LUMO gap), which determines the electronic properties of a polymer, can be calculated using DFT. These calculations provide valuable insights into the design rules for creating molecules with specific properties such as high charge mobility and stability. Understanding these properties is vital for the development of materials that perform well in real-world applications.

Due to the inverse relationship between accuracy and computational complexity, accurate QM methods are computationally intensive, especially for large systems or extensive chemical spaces.[53] This limitation requires the development of more efficient computational techniques. Semi-empirical QM methods, like GFN2,[54] offer a viable alternative. These methods are significantly faster than DFT, albeit with some loss in accuracy. GFN2, for example, has been shown to provide reliable geometries and approximate electronic properties, making it suitable for high-throughput screening of large molecular datasets. This trade-off between speed and accuracy is often acceptable in the early stages of material discovery, where rapid screening is essential to narrow down potential candidates.

The integration of QM methods, including both DFT and semi-empirical methods like GFN2, with machine learning is a key aspect of the approach presented in this thesis. By combining the detailed insights from QM methods with the predictive power of machine learning models, researchers can achieve a more efficient and accurate prediction of material

properties. This hybrid approach leverages the strengths of both techniques, facilitating the discovery of new materials with optimal properties for specific applications.

1.5 Machine Learning for Molecular Design

Machine learning (ML) has emerged as a powerful tool in the field of material science, offering the ability to predict material properties with high accuracy and efficiency. By training on large datasets of known materials, ML models can learn the complex relationships between molecular structures and their properties, enabling the rapid screening of vast chemical spaces. This capability is particularly important for exploring the vast chemical space of potential polymer structures and drug-like molecules, where traditional experimental methods would be too time-consuming and costly.[21, 22, 23, 24, 25, 26, 27, 28, 29]

In the context of π-conjugated polymers and drug-like molecules, ML models can predict properties such as reorganization energy, bandgap, stability, and micro-pK_a values.[55, 45, 46, 47, 48, 49, 50] These predictions are based on features derived from molecular structures, such as geometrical descriptors and electronic properties calculated using QM methods. The combination of QM and ML allows for the accurate prediction of material properties with significantly reduced computational cost compared to QM calculations alone. This integration enables researchers to quickly identify promising candidates for further investigation, thereby accelerating the material discovery process.

One of the key advantages of ML in molecular design is its ability to handle large and diverse datasets. ML models can be trained on a wide variety of data, including experimental results as well as computational predictions. This capability allows them to generalize from existing data to predict properties of new, unseen compounds. This generalization is crucial for discovering materials with novel properties that are not present in the training datasets. Moreover, advanced ML techniques such as neural networks, decision trees, and graph neural networks can capture non-linear relationships between features, enhancing the predictive accuracy of the models.

The integration of ML models with high-throughput experimental techniques could lead

to a more seamless and efficient discovery process, where computational predictions are rapidly validated and refined through experiments. Additionally, techniques such as transfer learning, where models trained on one type of data are adapted to another, and active learning, where models iteratively query new data points to improve performance, hold great promise in this regard. These advances in ML algorithms will further enhance their applicability in material science, making them indispensable tools for modern researchers.

1.6 Genetic Algorithms for Chemical Space Exploration

Genetic algorithms (GAs) are optimization techniques inspired by the process of natural selection.[56] They are particularly well-suited for exploring large and complex chemical spaces, where the goal is to identify structures with optimal properties. GAs use a population of candidate molecules that evolve over generations through selection, crossover, and mutation (Figure 1.3) to converge to a population with optimized properties. This evolutionary approach mimics natural selection, where only the fittest solutions survive and propagate, leading to a gradual improvement of the population over generations.

In the context of π-conjugated polymers, GAs can be used to explore the chemical space of different monomer combinations. By encoding the polymer structures as pairs of monomers, GAs can efficiently search for polymers with desired electronic and chemical properties. The fitness of each candidate molecule is evaluated based on its predicted properties, such as the HOMO-LUMO gap, guiding the evolution toward optimal solutions. This process allows researchers to systematically explore a vast number of potential structures and identify the most promising candidates for further investigation.

The effectiveness of this approach was demonstrated through a case study on biradical π-conjugated polymers.[57, 58] This case study illustrated how GAs could be used to navigate the chemical space and identify high-performing materials that meet specific criteria. The ability to combine multiple computational techniques into a cohesive workflow is one of the strengths of this integrated approach, offering a powerful toolkit for computational and material scientists. The success of these case studies underscores the potential for computational

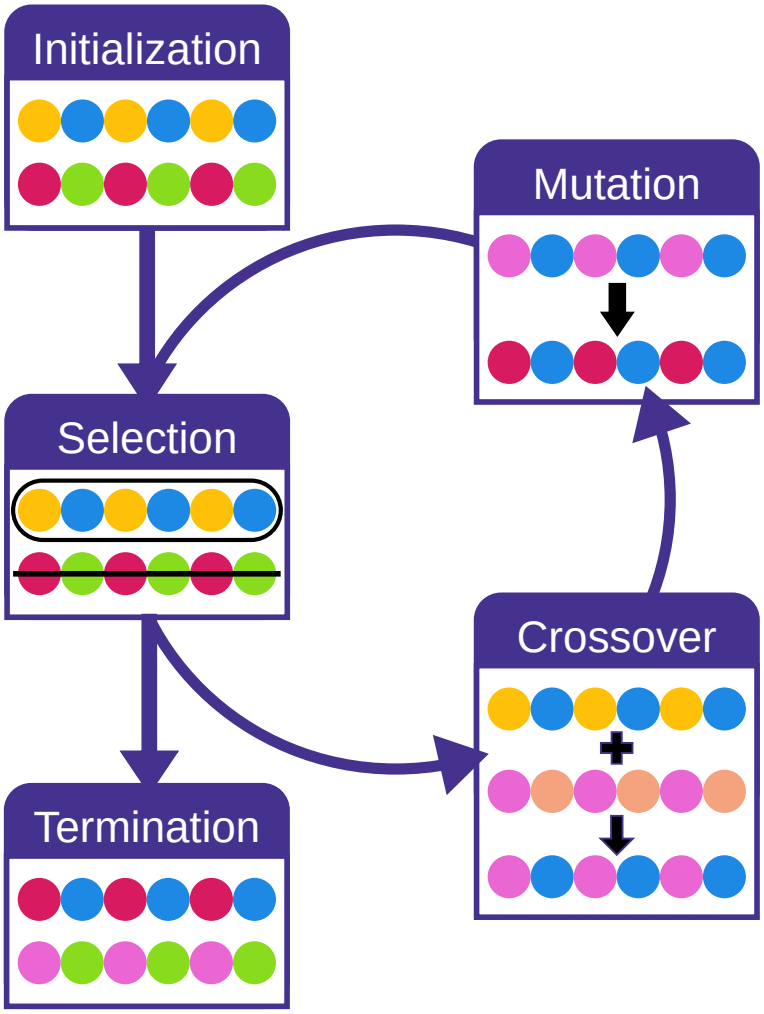


Figure 1.3: The five steps of the genetic algorithm - initialization, selection, crossover, mutation, and termination. The selection-crossover-mutation cycle is repeated a set number of times before stopping at the termination step.

methods to revolutionize material discovery.

1.7 Dissertation Overview

The integration of quantum mechanical methods, machine learning, and genetic algorithms represents a powerful approach to material discovery. Each of these techniques brings unique strengths: QM methods provide accurate property calculations, ML models offer efficient predictions, and GAs enable the exploration of large chemical spaces. Together, they form a comprehensive framework for the design and discovery of new materials. This integrated approach not only enhances the efficiency of the discovery process, but also opens up new possibilities for designing materials with tailored properties.

This thesis demonstrates the effectiveness of this integrated approach through several case studies on π-conjugated polymers and drug-like molecules. By leveraging the strengths of each technique, the proposed methodology can identify polymers with optimal electronic properties and molecules with favorable micro-pK_a values for specific applications. The case studies illustrate the practical implementation of the approach, highlighting the potential for accelerating the discovery of new materials. The successful application of these techniques in real-world scenarios underscores their practicality and effectiveness.

In conclusion, the combination of QM calculations together with ML and GA offers a promising pathway for material discovery. This integrated approach not only enhances the efficiency of the discovery process, but also opens up new possibilities for designing materials with tailored properties. As computational techniques continue to advance, the potential for discovering novel materials will only increase, paving the way for innovations in various fields of science and technology. The methodologies developed in this thesis provide a robust framework for future research, offering a roadmap for the continued exploration and optimization of materials with desirable properties.

Looking ahead, the integration of these computational techniques with experimental validation will be crucial for translating theoretical discoveries into practical applications. By creating a feedback loop between computational predictions and experimental results, re-

searchers can continuously refine their models and improve the accuracy of their predictions. This iterative process will accelerate the pace of material discovery and enable the development of next-generation materials with unprecedented performance and functionality.

Overall, the work presented in this thesis highlights the transformative potential of integrating advanced computational techniques in material science. The methodologies and case studies discussed provide valuable insights and practical guidance for researchers aiming to explore and design new materials. As the field continues to evolve, the integration of QM, ML, and GAs will play an increasingly important role in shaping the future of material discovery and innovation.

2.0 Machine Learning to Accelerate Screening for Marcus Reorganization Energies

This chapter is adapted from:

Omri D. Abarbanel, Geoffrey R. Hutchison; Machine learning to accelerate screening for Marcus reorganization energies. *The Journal of Chemical Physics* 7 August 2021; 155 (5): 054106. DOI: doi.org/10.1063/5.0059682.

It is a collaborative effort in which the author implemented the machine learning models, performed the calculations and data analysis, generated the figures, and wrote the manuscript; G.R.H. conceived and directed the project.

2.1 Summary

Understanding and predicting the charge transport properties of π-conjugated materials is an important challenge for designing new organic electronic devices, including solar cells, plastic transistors, light-emitting devices, and chemical sensors. A key component of the hopping mechanism of charge transfer in these materials is the Marcus reorganization energy, which serves as an activation barrier to hole or electron transfer. While modern density functional methods have proven to accurately predict trends in intramolecular reorganization energy, such calculations are computationally expensive. In this work, we outline active machine learning methods to predict computed intramolecular reorganization energies of a wide range of polythiophenes and their use towards screening new compounds with low internal reorganization energies. Our models have an overall root mean square error of $\pm$0.113 eV but a much smaller RMSE of only $\pm$0.036 eV on the new screening set. Since the larger error derives from high-reorganization energy compounds, the new method is highly effective to screen for compounds with potentially efficient charge transport parameters.

2.2 Introduction

Polythiophenes are a class of π-conjugated conductive and semi-conductive organic materials which can be used in many electronic devices, such as field-effect transistors[34, 35], organic solar cells[7, 8, 9], chemical sensors[10, 11, 12, 13, 14], and more[15]. The electronic properties of polythiophenes can be tuned across a wide range by various synthetic substitutions of the parent thiophene ring, which has enabled both fundamental studies and many applications.

The vast majority of polythiophene derivatives are p-type, with the charge transfer mediated by a hole transfer process[36, 37, 38]. Marcus-Hush charge transfer theory shows that the internal reorganization energy (λ), which describes the energy change required to distort geometry upon a charge transfer, is one important factor in the charge transfer rate and resulting charge mobility[29, 28, 36, 37, 38].

The internal reorganization energy λ of a molecule undergoing hole transfer to the same species can be calculated from four energies - the energy of the neutral molecule in the lowest energy geometry (E_0, "Neutral"), the energy of the cation at its lowest energy geometry (E_+, "Cation"), the energy of the cation at the geometry of the neutral species (E_+^*, "Cation@Neutral"), and the energy of the neutral molecule at the geometry of the cation (E_0^*, "Neutral@Cation")[36, 38] (Figure 2.1). The λ can then be calculated from those energies according to the following formula 2.1:

$$\lambda = \lambda_0 + \lambda_+ = (E_0^* - E_0) + (E_+^* - E_+) \tag{2}$$

Calculating the internal λ of polythiophenes using density functional theory (DFT) calculations requires two geometry optimizations (of both the neutral and cationic species) and can be computationally expensive, as the calculation time increases with the length the polythiophene chain. Recent work on the approximate density functional GFN2 method[54] has shown accurate geometries and excellent correlation with coupled-cluster methods for conformers[53]. We attempted to correlate reorganization energies computed with GFN2 with those computed with the B3LYP DFT method[59, 60]. As discussed below, no significant correlation was found.

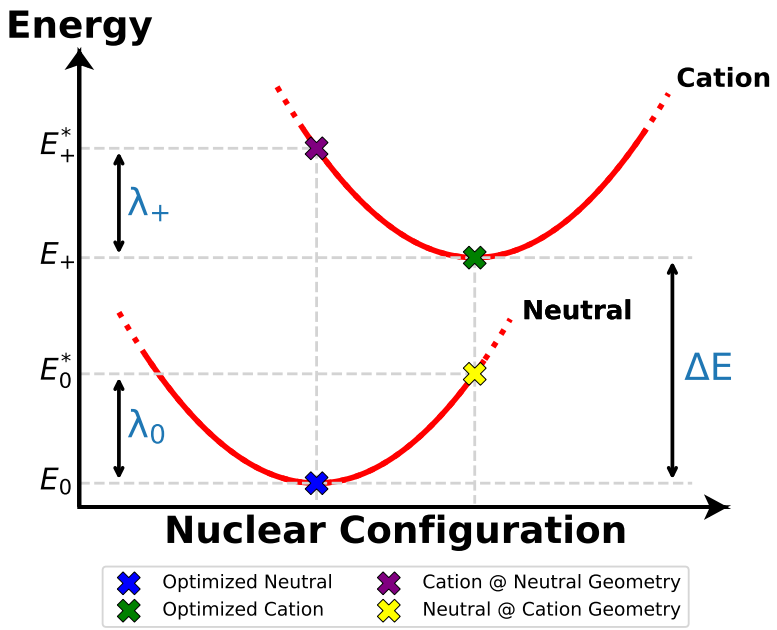


Figure 2.1: Internal reorganization energy for hole transfer.

In this work, we instead focus on predicting internal reorganization energies λ using machine learning (ML) methods, using a minimal amount of B3LYP-calculated λ as a training set.

In recent years ML has been applied widely, with a goal of accelerating quantum chemical calculations that would otherwise have large computational costs. Calculating electronic properties with traditional methods can be computationally expensive and take between hours to weeks to finish, depending on the size of the system and the type of calculation. ML has shown a great potential in calculating electronic structure properties, drug discovery, materials research, and more[21, 22, 23, 24, 25, 26, 27, 28, 29]. Training a ML model is also time consuming as well, since it requires a large data set for training and finding an accurate ML method and representation for that specific application can be exhaustive, but once a model has been properly trained, evaluation for new calculations can be performed in seconds or less.

In this work, we have developed a machine learning filter for predicting the internal

reorganization energy of organic electronic materials. At present, we find the accuracy to be greater for compounds with low reorganization energy - as such it proves more useful for ignoring compounds expected to have high barriers for charge transport than as a fully accurate surrogate across the entire range considered. Nevertheless, since key applications require efficient charge transport, and thus low reorganization energies, we demonstrate its use in efficiently screening a pool of possible co-polymers. Finally, we discuss frequent chemical motifs among compounds with low predicted reorganization energies.

2.3 Methods

2.3.1 Computational Methods

Input files for each oligomer were created by combining the corresponding SMILES strings of its monomers and using OpenBabel version 3.1.0[61] to generate a 3D geometry.[62] All GFN2 calculations were performed using xTB version 6.0 [54]. All DFT calculations were done using the B3LYP functional[59, 60] with the 6-31G* basis set,[63] calculated with Gaussian 09,[64] for comparison with previously published internal reorganization energies.[29, 36]

Random forest, gradient boosted trees and kernel ridge regression models were implemented using Scikit-Learn version 0.20.0[65]. Neural network model was implemented using Keras version 2.3.1[66] with TensorFlow version 2.1.0[67] backend.

The data that support the findings of this study, including SMILES for all monomers, all Python code and notebooks are openly available at `https://github.com/hutchisonlab/ReorganizationEnergy`

2.3.2 Data Set

Our data set derives from 253 thiophene-based monomers. The monomers have different functional groups at the 3 and 4 positions, while connected to other monomers at the 2 and 5 positions (Figure 2.2), yielding a total of 31,878 possible copolymers, plus 253 homopolymers, to a total of 32,131 possible oligomer families. We used our available monomers to create a

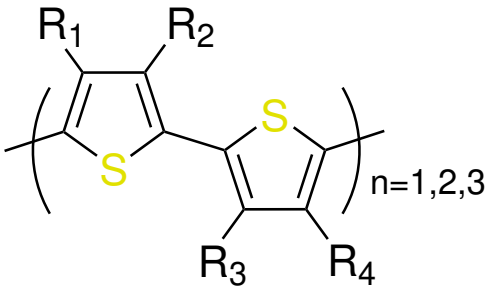


Figure 2.2: Thiophene based oligomers, length of 2, 4, and 6 monomers - named dimers, tetramers, and hexamers, respectively.

list of possible oligomers made from two, four, and six monomers - dimers, tetramers, and hexamers, respectively (Figure 2.2).

Calculating the λ of long oligomer chains using traditional DFT methods can be time-consuming and computationally expensive. However, previous studies have claimed that six-membered oligomer chains can closely estimate the λ of longer chains[36]. However, quantum mechanical calculations, especially when optimizing molecular geometries, drastically increase with the length of the oligomer (Figure 2.3a). We therefore have explored different ways to minimize the calculation time, such as using an approximate method, GFN2, and using shorter oligomers.

At first, we calculated the λ of all the oligomers using GFN2-xTB, an approximate density functional tight-binding method developed by the Grimme group[54], to see if it can be used as an accurate surrogate for B3LYP-computed reorganization energies. This method produces accurate geometries and is considerably faster than B3LYP calculations (Figures S7 and 2.3a). However, the λ calculated using GFN2-xTB does not correlate well with the λ calculated using B3LYP (Figure S2).

In contrast, while the energies have little correlation, we have found that the geometries of both the neutral and cation species calculated using GFN2-xTB have a significant correlation with those calculated using B3LYP — specifically, the average dihedral angle and the average inter-ring bond length. Therefore, instead of using GFN2-xTB to calculate the

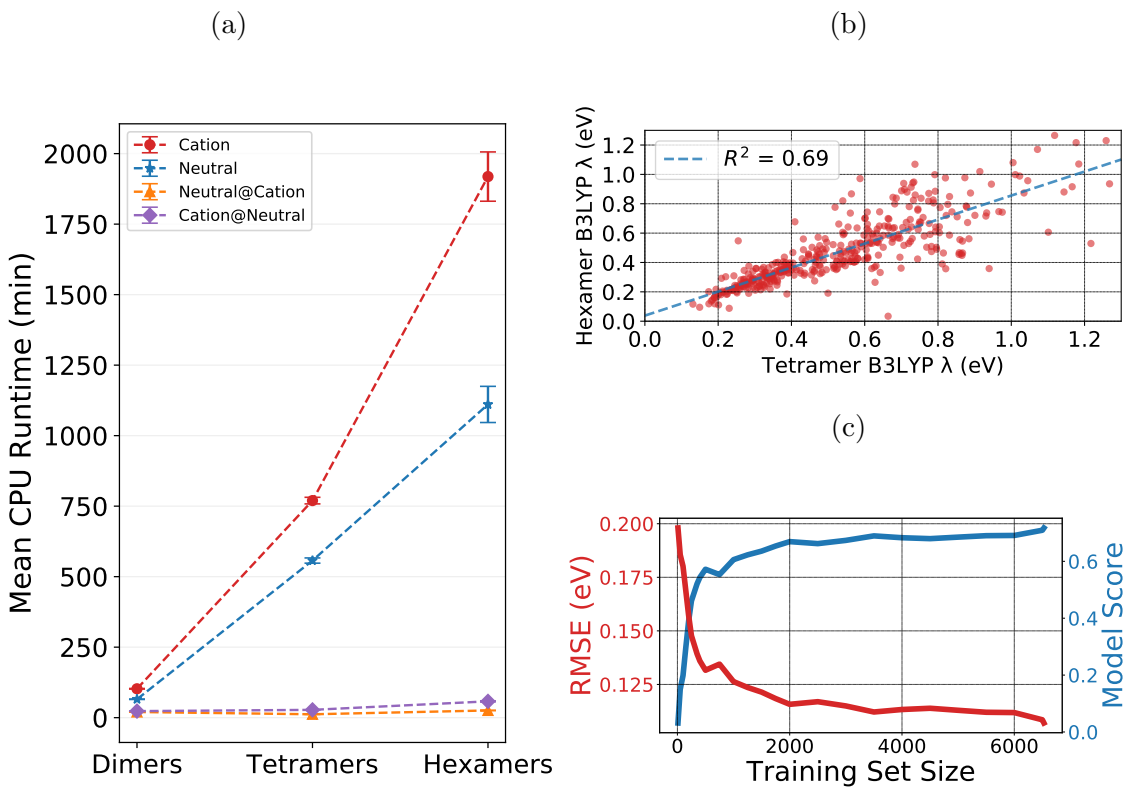


Figure 2.3: **(a)** Mean CPU run time of the 4 different calculations using B3LYP, **(b)** Correlation between tetramers λ and hexamers λ, trendline indicated robust linear regression fit, **(c)** training set size effect on the RF model score.

λ, we considered using ML methods using the geometrical descriptors obtained from the GFN2-xTB calculations. Likely, while the geometric minima correlate well between GFN2 and B3LYP, the shape of the potential energy surfaces differ substantially away from the local minima.

In addition, we considered a correlation of λ between shorter and longer oligomers (Figure S1), since shorter oligomers are faster to optimize. We did not find such a correlation between the B3LYP-computed λ of the dimer and tetramers, or the dimers and hexamers. We did find, however, a good correlation between the tetramers and hexamers (Figure 2.3b). Thus,

to develop an adequate training set, more tetramers than hexamers can be used, considerably reducing the calculation time. We therefore used a data set made up of mainly tetramers plus a small number of hexamers. We increased the training set in batches until we saw no significant improvement in the model score (Figure 2.3c) and decrease in RMSE. Our final data set consisted of 7020 tetramers and 408 hexamers with B3LYP-calculated λ between 0 and 2 eV. We chose this range as we assumed that oligomers with λ larger than 2 eV are irrelevant to our study.

2.3.3 Representation and Model Selection

For the representation of the oligomers in the ML model we began with the monomer ID and the oligomer length. In addition, we added the average dihedral angle between the monomers of each oligomer and the average inter-ring bond length of each oligomer, for both the neutral and cation species of each oligomer, as calculated using GFN2. We saw correlation (e.g., R^2 between 0.57 and 0.74) between those geometric values calculated with GFN2 and with B3LYP (Figure S3). Using those starting features gave us decent preliminary results. Next, we added an extended circular finger print (ECFP4) 2048 bit representation[68] using RDKit[69], which increased the R^2 and decreased the RMSE of the model significantly, likely by describing local functional group effects on reorganization energies.[29] The final step was to add a new feature to represent the size of the π-system in the oligomers (Figure S10), as we hypothesized that a highly-conjugated oligomer will contribute to a lower λ. Adding this final feature moderately improved the model (Table 2.1).

Using our best representation, we trained five different ML models with our data set: a random forest model, a gradient boosting trees model, a ridge regression model, a kernel ridge regression model and a neural network model. The first two are ensemble methods based on decision trees, which combine several weighted trees into one model. The random forest model builds a large number of random sets of decision trees[70], hence the name, while the gradient boosted trees model builds nested decision tree one at a time, improving over the previous tree[71]. Ridge regression fits the given data into a function in a way that minimizes the coefficients of the parameters by penalizing the cost function with the square

Table 2.1: Cross-validated R^2 and RMSE, averages and standard errors, to show model development improvement as new features are added to the representation.

	Geometrical Data		Geometrical Data + π-System Size		ECFP4		Geometrical Data + ECFP4		Geometrical Data + ECFP4 + π-System Size	
	R^2	RMSE (eV)	R^2	RMSE (eV)	R^2	RMSE (eV)	R^2	RMSE (eV)	R^2	RMSE (eV)
Run 1	0.533	0.138	0.536	0.138	0.692	0.112	0.706	0.109	0.719	0.107
Run 2	0.526	0.140	0.548	0.137	0.653	0.120	0.677	0.116	0.681	0.115
Run 3	0.521	0.140	0.559	0.134	0.654	0.120	0.661	0.118	0.663	0.118
Average	**0.526** ± **0.003**	**0.139** ± **0.001**	**0.548** ± **0.007**	**0.136** ± **0.001**	**0.666** ± **0.013**	**0.117** ± **0.003**	**0.681** ± **0.013**	**0.114** ± **0.003**	**0.688** ± **0.017**	**0.113** ± **0.003**

sum of the coefficients times the regularization parameter alpha[72], and the Kernel Ridge Regression which works similar to the Ridge Regression but with an addition of a kernel trick which allows to fit a non-linear function[73]. The fourth ML model, a neural network, has been widely used in many classification and regression applications. Neural networks are built in layers, where each node in each layer is connected to all the nodes in the next layer. A mathematical loss function is dictating how much each node is contributing to the network, creating a complex structure that can predict values or classify objects[74].

In a random forest model, the key hyperparameter is only the number of trees in the forest. The greater the number of trees is likely to yield better predictions but also increases the time it takes to train. Moreover, the model eventually reaches a prediction ceiling where adding more trees will not improve the model. We optimized the number of trees in the random forest model, ranging from 10 to 1500, and recorded the training time, the Scikit-Learn built-in *score* function value for random forest models, which is comparable to the coefficient of determination, R^2, and the root mean square error (RMSE) for to the test set. As indicated in Figure S11, 50 trees are the optimal number for the random forest, as it gives the optimal training time, of about 25 seconds, while having the highest score and lowest RMSE.

For the gradient boosting trees model, there are several hyperparameters to optimize — including the number of trees, maximum tree depth, minimum sample split, learning rate,

Table 2.2: R^2 and RMSE results for three runs for each machine learning method used. The training and test sets in each run were the same for each method. Averages are presented with standard error.

	Random Forest		Neural Network		Gradient Boosting Trees		Ridge Regression		Kernel Ridge Regression	
	R^2	RMSE (eV)	R^2	RMSE (eV)	R^2	RMSE (eV)	R^2	RMSE (eV)	R^2	RMSE (eV)
Run 1	0.716	0.108	0.631	0.122	0.645	0.121	0.655	0.118	0.683	0.113
Run 2	0.662	0.118	0.653	0.120	0.575	0.134	0.644	0.121	0.680	0.115
Run 3	0.685	0.114	0.623	0.125	0.612	0.129	0.666	0.117	0.686	0.114
Average	**0.687** ± **0.016**	**0.113** ± **0.003**	**0.636** ± **0.009**	**0.122** ± **0.001**	**0.611** ± **0.020**	**0.128** ± **0.004**	**0.655** ± **0.006**	**0.119** ± **0.001**	**0.683** ± **0.002**	**0.114** ± **0.001**

and the loss function. Optimization started using the common starting parameters of 1000 trees, unlimited maximum tree depth, minimum sample split of 2, learning rate of 0.01, and the *least squares* loss function. Parameters were manually sampled, comparing the mean square error (MSE) score. This initial sampling did not noticeably affect the performance relative to the random forest and the neural network model. Therefore a more exhaustive grid search over these hyperparameters was not performed.

For the Ridge regression model there is only one hyperparameter to optimize, alpha, which was set to 9 after scanning over a range of values and finding the one with the highest score. Similarly, for the Kernel Ridge Regression model we tested linear and polynomial kernels and found that a third-degree polynomial kernel gives the best result. The alpha parameter was set to 60 after optimization by scanning over a range of values and finding the one that gave the highest score.

As for the neural network model, we used Bayesian optimization using the HyperOpt and Hyperas Python packages [75, 76], to find the optimal number of hidden layers, the number of nodes in each layer, and the dropout amount. We searched over a space of 1 to 3 hidden layers, 20 to 200 nodes per layer, and dropout between 0 to 0.5. We found that 2 hidden layers, size 127 and 109 nodes respectively, and the dropout amount of 0.005 for the first hidden layer and 0.448 for the second hidden layer, are the optimal values for this neural network. We used the Continuously Differentiable Exponential Linear Units

(CELU) activation function[77], as implemented in the EchoAI Python package[78], for the input and both hidden layers, as it outperformed other functions — including the widely used Rectified Linear Unit (ReLU) function. The output layer consists of a single node as standard for regression, with a linear activation function. Training was performed using the Adam optimizer[79] using the mean square error (MSE) loss function. In order to reduce the number of hyperparameters needed to optimize, the *ReduceLROnPlateau* and the *EarlyStopping* Keras functions[66] were used to tune the learning rate during the training of the model and stop the training once there is no further improvement. This effectively optimized the learning rate and the number of epochs for the neural network training.

For cross-validation, each model was trained on three different train-test split sets, using the Scikit-learn *train_test_split* function[65] using random state values of 0, 42, and 420. We saw that the random forest model outperformed all other models in both the R^2 value and the root mean square error (RMSE) (Table 2.2). We therefore used the random forest model as our model of choice for the remaining work.

2.4 Results and Discussion

In order to see how well the final random forest model can predict the λ of unseen oligomers we split the data set into 85%-15% train-test sets, respectively, comparing the trained model prediction of λ of the test set to the B3LYP calculated λ. The correlation graph between the predicted and calculated energies (Figure 2.4) shows good correlation with unitless coefficient of determination, $R^2 = 0.717$ and root mean square error, $RMSE = 0.105$eV for the tetramers, $R^2 = 0.737$ and $RMSE = 0.140$eV for the hexamers, and $R^2 = 0.719$ and $RMSE = 0.107$eV in total.

Moreover, as Figure 2.4 shows, the correlation also exhibits heteroscedasticity, where there is better correlation for oligomers with lower λ and worse for compounds with greater reorganization energies. This shows that predicting the λ for oligomers with geometric differences between the neutral and cation species is a complex task. In all likelihood there are many possible geometric changes between neutral and cation geometries, and as such

the limited training set makes it challenging for the model to properly account for all reorganization in compounds with large λ. Similar heteroscedastic behavior can be seen in the correlation between the tetramers and hexamers (Figure 2.3b. In principle, some of the heteroscedasticity in the predictions could be reduced by using more, or even only, hexamers in the training of the model. However, calculating the B3LYP λ for hexamers is computationally expensive — which runs counter to the benefit of the ML model as a surrogate for the calculations.

For screening, where the intent is to find candidates with low λ, the larger heteroscedastic error for higher λ compounds has only a small effect — there is better correlation for compounds with small internal reorganization energies. Therefore we can use the random forest model as a first, rapid screening tool to find oligomers with low λ.

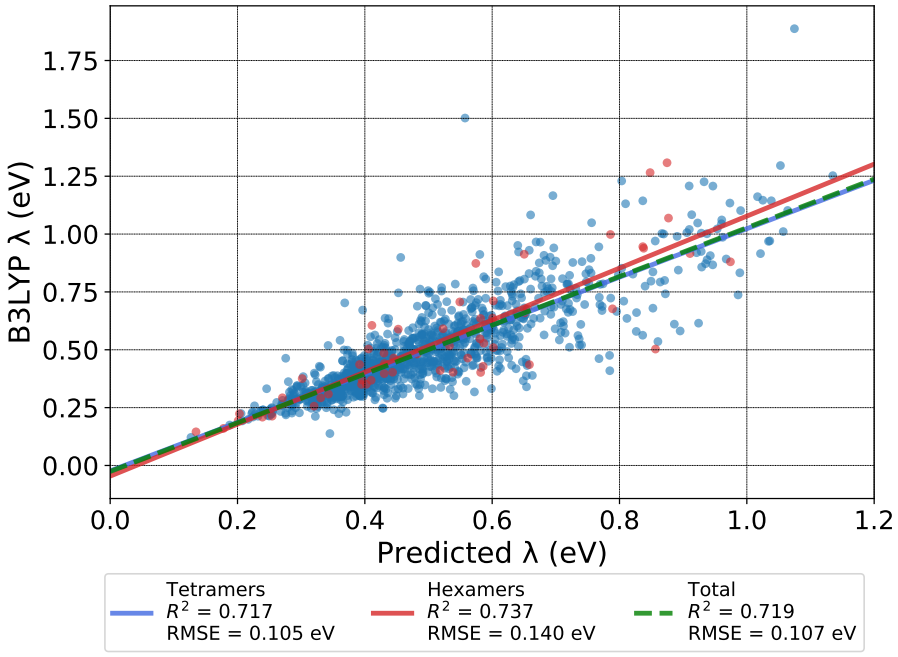


Figure 2.4: Correlation plot between the random-forest predicted λ and the B3LYP calculated λ for the tetramers and hexamers in the test set.

The random forest regression model, as implemented in the scikit-learn package, has a *feature_importance_* function[65], which enables exploration of the features in the representation that contribute the most to the model (Figure S12). It is clear that the most

important feature is the average inter-ring bond length of the neutral oligomer, as calculated using xTB-GFN2. While the bond lengths are expected to change going from the neutral to cation geometries, this is a surprising effect as the correlation between the neutral inter-ring bond length and the B3LYP-calculated λ is weak ($R^2 = 0.233$, Figure S14). Much like the overall reorganization energies, the correlation between GFN2-computed and B3LYP-computed inter-ring bond lengths shows the same heteroscedasticity, which may explain some of the feature importance. The second most important feature is the π-system size descriptor, which agrees with the hypotheses that bigger π-conjugated systems promote lower λ. The third most important feature is the ECFP bit number 1019, which indicates the existence of an sp^3 hybridized carbon in the oligomer (Figure S13). Two possible explanations exist for this feature — that a sp^3 hybridized carbon breaks conjugation and as discussed below, the CH_2 group may promote a less planar conformation. Interestingly, the monomer numbers, although used as a categorical feature with a seemingly arbitrary assignment meant for naming only, do appear to contribute to the model as the fourth and sixth most important features. The rest of the features are the other geometrical information we encoded into the representation, followed by the rest of the ECFP bits which minimally contribute to the model.

After using 85% of the training set to train the model for testing purposes, the final random forest model was trained using the full data set for screening a larger validation set to predict the λ of 24,853 tetramers and 31,722 hexamers that were not part of the original data set. From those new predicted λ, oligomers with $\lambda < 0.3$ eV were filtered to compute the full B3LYP λ, including 660 tetramers and 1753 hexamers with low λ (Figure 2.5 a, b). The increase in the number of oligomers with $\lambda < 0.3$ev from the tetramers to the hexamers agrees with assessment of the inverse relationship between the length of the oligomer and its reorganization energy[36].

We also looked to trends in the predictions in order to see if there are monomers that repeatedly contribute to oligomers with low λ (Figure 2.5 c, d). For both tetramers and hexamers, the monomers number 47, 110, 158, 213, 258, and 283 are found frequently (Figure 2.6). As it can be seen, all the best performing monomers have a fused aromatic system on the thiophene backbone, supporting our hypothesis that a larger π-system contributes to

low λ. Moreover, excluding monomer 253 which only has one, all of the monomers have two aromatic nitrogen atoms in the 3- and 4- positions on the thiophene ring. We hypothesize that steric considerations contribute, as CH_2 groups in these positions increase the steric repulsion between neighboring monomers forcing a non-planar, twisted chain conformation and increasing λ [36]. In addition to that, monomers with similar motifs, such as nitrogen atoms in the 3- and 4- positions, and especially a thiadiazol group, have been shown to be pro-quinoidal monomers which has some quinoidal character instead of an aromatic one[80, 81, 82, 83]. This behavior contributes to a higher double bond character of the inter-ring bond, decreasing its length and restricting the rotation of the dihedral angle. This restrains the conformational change the oligomer undergoes upon a hole transfer, which contributes to a low λ.

In order to validate the accuracy of the full random forest model the 300 tetramers and 150 hexamers with the lowest predicted λ were selected and the B3LYP λ was computed to compare with the random forest model prediction (Figure 2.7). While the predicted values are not perfect, the low RMSE (0.036 eV) of the prediction versus the calculated λ, indicates that the model is robust and accurate at this new validation set and thus can be used as a first step in finding conjugated materials with better charge transport properties. Interestingly, of the 50 hexamers with the lowest B3LYP λ, 44 oligomers had monomer 47 as one of their monomers, and the hexamer with the lowest B3LYP λ consists of the homo-oligomer of monomer 47, with $\lambda = 0.051$ eV (Figure 2.7b). This fragment, and related monomers, is frequently used in top organic photovoltaic materials.

Moreover, the dihedral angle between the best performing hexamers is close to 180(Table S1), or in other words - flat, and is only minimally changing between the neutral and cation species (Table 2.3). This further strengthens the hypothesis that in addition to a large π-system, better conjugation and planar chain conformations contribute to the low λ. In addition to the dihedral angle, the best performing hexamers exhibit a minimal change between the neutral and the cation bond lengths.

For comparison, in a recently published paper the Atalay group have used deep neural network and kernel ridge regression models and were able to predict the reorganization energy with a high precision compared to the DFT calculated values[29]. In contrasts, while their

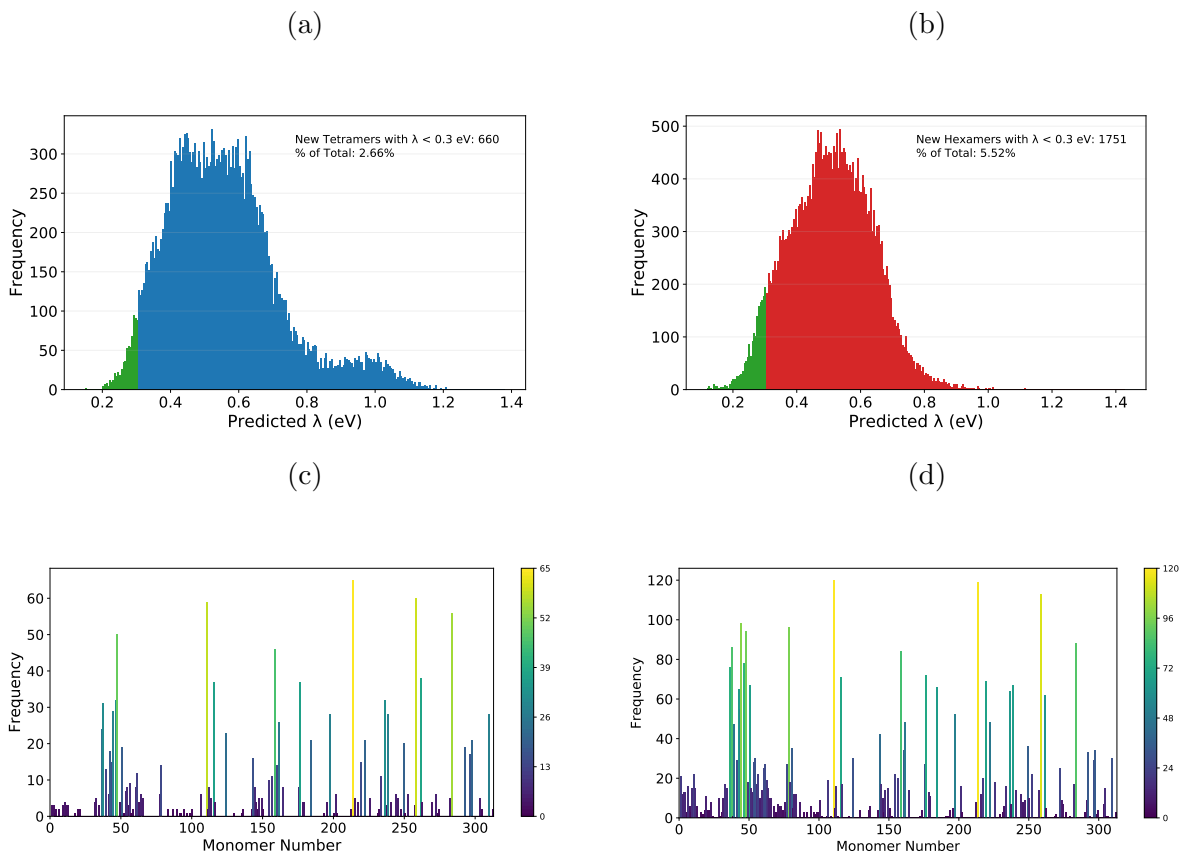


Figure 2.5: λ predictions for new tetramers and hexamers. Histogram of predicted λ for the new tetramers **(a)** and hexamers **(b)**, with tetramers and hexamers with $\lambda < 0.3$ eV colored in green. Histograms of the common monomers for the tetramers **(c)** and hexamers **(d)** with $\lambda < 0.3$ eV.

data set consists of ring-fused conjugated molecules, which are mostly planer and rigid, out data set is made out of long, flexible, oligomeric chains. While they used similar molecular representation in their model, it does not capture any molecular deformation that happen upon a hole transfer. However, unlike the data set we have used here, their molecules present a minimal conformational difference between the neutral and cation species, which can explain the high accuracy and precision of their model. While our model is unable to predict the λ of all oligomers with low accuracy, we have shown that it can be used

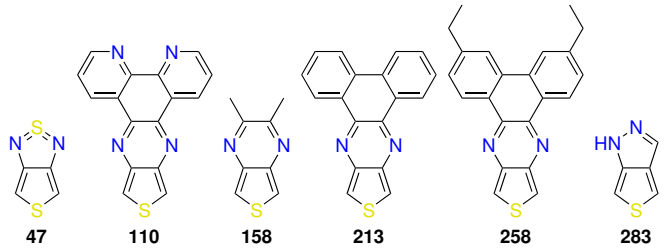


Figure 2.6: The top 6 common monomers in oligomers with predicted $\lambda < 0.3$ eV.

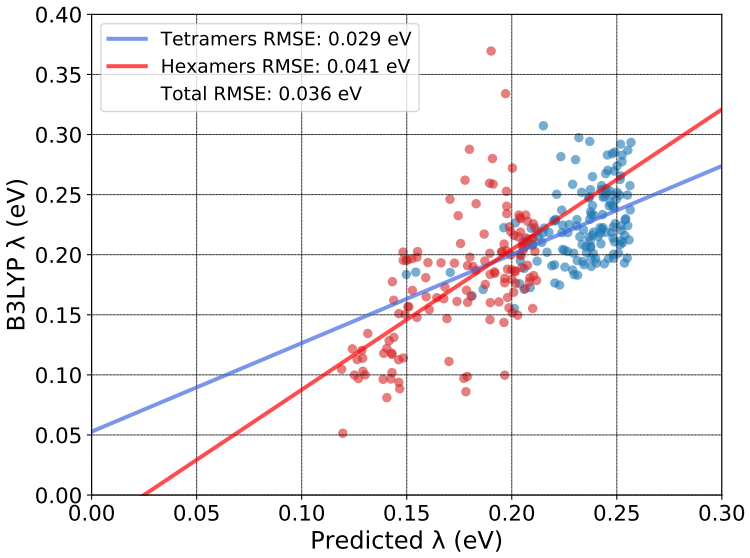
successfully as a screening tool to separate oligomers with low λ from the rest.

2.5 Conclusions

In this work we have shown that a random forest model can be used as a rapid screening tool to find thiophene-based oligomers with low and high λ. Our goal was to train a model by minimizing the calculation time required to generate the training set by calculating the λ of shorter oligomers (i.e., tetramers), correlating with the λ of longer lengths. The resulting random forest regression model can predict thousands of new oligomers in seconds, yielding a list of potential oligomers with low λ for further screening. The model has an overall RMSE of $\pm 0.113eV$ but a much smaller error of $\pm 0.036eV$ on this validation set of low-reorganization energy targets, highlighting the utility in computational screening. Comparing the time required to generate the test and validation sets to the possible time required to calculate all 31,878 tetramers and 31,878 hexamers, the RF model yields a $\sim$13$\times$ speedup. If the model were used across a larger search space, larger speedups would likely result. We intend to use the model in future computational screening efforts.[84]

From the predictions of the model and the relative feature importance, it is clear that oligomers with large, conjugated π-systems have lower internal reorganization energies. In addition to a large π-system size, monomers with low steric bulk, which minimally change conformation upon a hole transfer, that also yield a high degree of delocalization and π

(a)



(b)

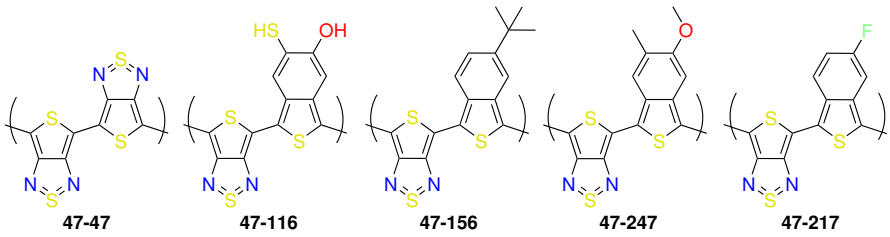


Figure 2.7: **(a)** Correlation plot for the tetramers and hexamers with low λ where the trendline indicate robust linear regression fit, the tetramers are in blue, and the hexamers are in red. **(b)** The top 5 hexamers with the lowest B3LYP calculated λ. The numbers represent the two monomers in the chain.

Table 2.3: The monomer numbers, the predicted and calculated B3LYP λ, The GFN2 and B3LYP geometrical data of the average change in dihedral angles between the neutral and cation species, and average change in the inter-ring bond length of both neutral and cation species for the five hexamers with the lowest B3LYP λ.

Monomer 1	Monomer 2	Predicted λ	B3LYP λ	GFN2 ΔAngle (	B3LYP ΔAngle (	GFN2 ΔBond Length (Å)	B3LYP ΔBond Length (Å)
47	47	0.120	0.051	0.063	0.005	0.004	0.005
47	116	0.141	0.081	0.034	0.097	0.005	0.008
47	156	0.178	0.086	0.415	0.019	0.006	0.010
47	247	0.147	0.088	0.352	1.628	0.005	0.010
47	217	0.146	0.094	0.032	4.460	0.006	0.012

orbital overlap between the monomers, also contributes to low λ. One monomer in particular, with a thiadiazole group, is frequently observed in compounds with low internal reorganization energy. Moreover, aromatic nitrogen substituents are frequently observed in such compounds. All the top oligomers also share similar geometries, i.e., being almost completely flat, and only exhibit minimal changes in geometry upon a hole transfer. Future work can consider a similar method for internal reorganization energies of n-type electron transfer or other calculated properties requiring multiple time-intensive computational steps. Future models should also address the potential for conformational entropy, since multiple low-energy conformers likely exist and can affect the reorganization energy.[85, 86]

3.0 Strategies for Computer-Aided Discovery of Novel Open-Shell Polymers

This chapter is adapted from:

Omri D. Abarbanel, Julisa Rozon, Geoffrey R. Hutchison; Strategies for Computer-Aided Discovery of Novel Open-Shell Polymers. *Journal of Physical Chemistry Letters* 2022, 13, 9, 2158–2164. DOI: doi.org/10.1021/acs.jpclett.2c00509.

It is a collaborative effort in which J.R. and the author performed the calculations; The author performed the data analysis, generated the figures, and wrote the manuscript; G.R.H. conceived and directed the project.

3.1 Summary

Organic π-conjugated polymers with a triplet ground state have been the focus of recent research for their interesting and unique electronic properties, arising from the presence of the two unpaired electrons. These compounds are usually built from alternating electron-donating and electron-accepting monomer pairs which lower the HOMO-LUMO gap and yield a triplet state instead of the typical singlet ground state. In this paper we use density functional theory calculations to explore the design rules that govern the creation of a ground state triplet conjugated polymer, and find that a small HOMO-LUMO gap in the singlet state is the best predictor for the existence of a triplet ground state, compared to previous use of pro-quinoidal bonding character. This work can accelerate the discovery of new stable triplet materials by reducing the computational resources needed for electronic-state calculations and the number of potential candidates for synthesis.

3.2 Introduction

Organic π-conjugated polymers have been a focus of fundamental research for many years thanks to their delocalized electronic properties, which can be used in a wide variety of applications[16, 17]. The ground state of the vast majority of those polymers is a singlet state (S_0), which can be excited to a triplet (T_1) state via different pathways, such as intersystem crossing or reverse-intersystem crossing (RISC)[87, 88, 89, 8, 90]. However, in recent years organic π-conjugated polymers with a triplet *ground state* (dubbed "T_0") have been discovered and studied for their unique electronic, optical, and magnetic properties arising from their unpaired electrons. Such ground-state triplet materials have found applications as varied as batteries[91], supercapacitors[92], non-linear optics[93], and many others[94, 95, 96, 83].

Understanding the design rules for the synthesis of such molecular diradicals can aid with the discovery of new materials. Those structure/function correlations can help us determine the type of monomers that will promote a triplet ground state, how the electronic structure is affected, and how to design new materials. For example, previous studies have used "pro-quinoidal" monomers in order to create high-spin polymers, suggesting that a quinoidal bonding character helps to stabilize the diradical polymer[80, 81, 83, 82]. However, others challenge this, by suggesting instead that instead, an aromatic bonding character stabilizes the diradical ground state[95]. These two opposing hypotheses can lead to different design rules, but by finding the best predictors of a stable ground-state triplet, we can assist in the discovery process.

In this work, we have used dispersion-corrected ωB97X-D3 density functional theory (DFT) method to calculate the ground state energies of both singlet and triplet states of various π-conjugated oligomers. Our data set consists of 11 donor monomers and 12 acceptor monomers (Figure 3.1), most previously studied by the Azoulay group[81, 97, 98, 99, 100, 101, 92], yielding a set of 132 oligomers. The geometry optimization steps for both singlet and triplet states, the single-point electronic energy calculations, and the calculation of ΔE_{T-S} were performed as described in the **Computational Methods** section below. Thus, for each oligomer, both singlet and triplet states were optimized to find the lowest energy geometry

and corresponding ground-state energy.

To find key predictors of ground state triplets, we compared the a variety of electronic and geometric properties as predictors of the energy difference between the triplet and singlet energies. In addition, we considered different strategies to favor triplet stabilization, such as changing the heteroatom in the polymer backbone, and forcing quinoidal bonding character. This work can aid in the discovery of new and novel open shell materials by increasing the search speed and decreasing the search space of potential candidates.

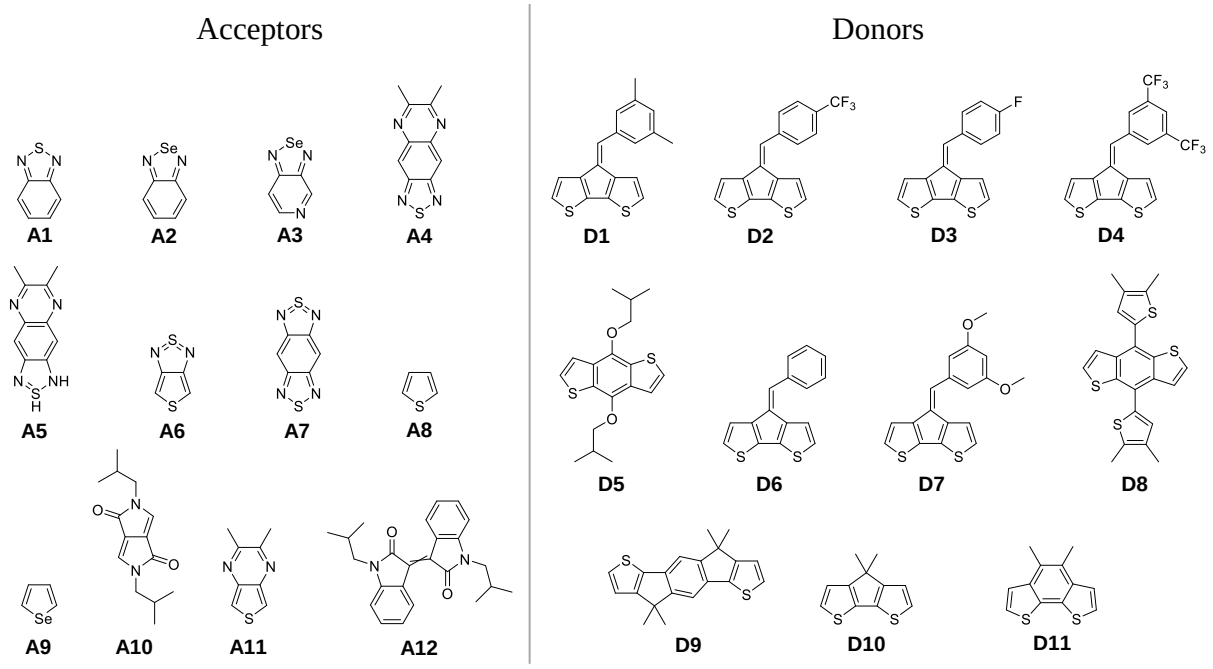


Figure 3.1: The acceptors and donors used to create the tetramers.

3.3 Results and discussion

3.3.1 Inter-Monomer Bond Length

As mentioned, one of the working hypotheses for the stability of a triplet ground-state is through a bi-radical system, in which each unpaired electron is in different singly-occupied

molecular orbitals (SUMO), and the formation of a semi-quinoidal bonding pattern in the polymer backbone[81, 80, 96]. This suggests that the bond between the monomers should have some double-bond character — and thus be shorter to stabilize the triplet ground-state. We measured the inter-monomer bond length of the oligomers, i.e. the bond between each donor and acceptor monomer (D-A-D-A-D-A-D-A), as a metric of quinoidal character, and compared this geometric measure to the difference between the electronic energies of the triplet and singlet species ΔE_{T-S} (Figure 3.2).

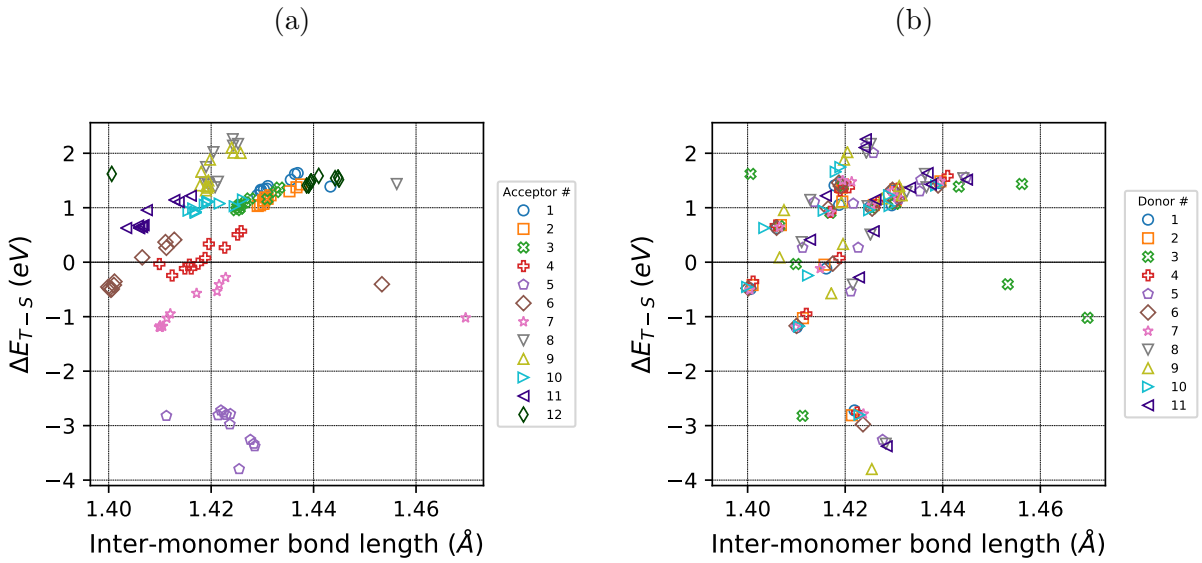


Figure 3.2: Correlation between ΔE_{T-S}, in eV, and the inter-monomer bond length, in Å, grouped by **a** acceptor number and **b** donor number.

Our calculations, however, show very little correlation between the inter-monomer bond length and the stability of the triplet ground state across all 132 oligomers. While there is a weak trend for some families of oligomers that share the same acceptor (Figure 3.2a), there are a few exceptions. In some cases, such as for the oligomer of donor D3 and acceptor A7, while it had a stable triple ground state - its calculated inter-monomer bond length was the longest at 1.47 Å. In contrast, the oligomer of donor D3 and A12 does not yield a stable triplet ground state, but has one of the shortest calculated inter-monomer bond lengths at

1.40 Å.

Furthermore, oligomers with acceptor A5 yield a negative slope between the inter-monomer bond length and ΔE_{T-S}, while all other families have a positive slope (Table B1). Thus, while looking at some specific acceptors we can see a trend, there is no overall correlation between the stability of the triplet ground state and the inter-monomer bond length. In addition, there is no correlation between the inter-monomer length and the donor number (Figure 3.2b).

While the pro-quinoidal design rule might work for specific cases, the stability of the triplet ground state instead comes from a broader property — narrowing of the singlet HOMO-LUMO gap. While it has been shown that one strategy to lower the HOMO-LUMO gap is by quinoidal bonding[102, 103], it is by far not the only design rule.

3.3.2 Triplet-Singlet Correlation

A small HOMO-LUMO gap has been shown to promote a lower triplet energy level, increasing the likelihood of a high-spin ground state as the frontier molecular orbitals (MO) become closer energetically [81, 104, 105]. We therefore compared the energy difference between the singlet and triplet of each oligomer (ΔE_{T-S}) versus the HOMO-LUMO gap of the corresponding singlet-state oligomer, both in eV (Figure 3.3). Figures 3.3a and 3.3b show the same correlation and only differ by grouping of the acceptor number and donor number, respectively. Out of 132 oligomers, 35 had ground-state triplet states, based on the optimized geometries and ωB97X-D3/def2-SVP single-point energies. A significant correlation between those energies can be seen, with a correlation of determination (R^2) of 0.96 and a linear relation with an x-axis intercept at 3.84 eV (Eq. S5).

As seen in Figure 3.3b, there is very little correlation between the donor identity and ΔE_{T-S}. However, from Figure 3.3a it can be seen that oligomers that share the same acceptor monomer are grouped in close proximity, indicating that a high-spin system is strongly dependant on the identity of the acceptor.

While a spin-polarized singlet state may be lower in energy than the restricted singlet considered here, the strong trend indicates that as the singlet HOMO-LUMO gap decreases,

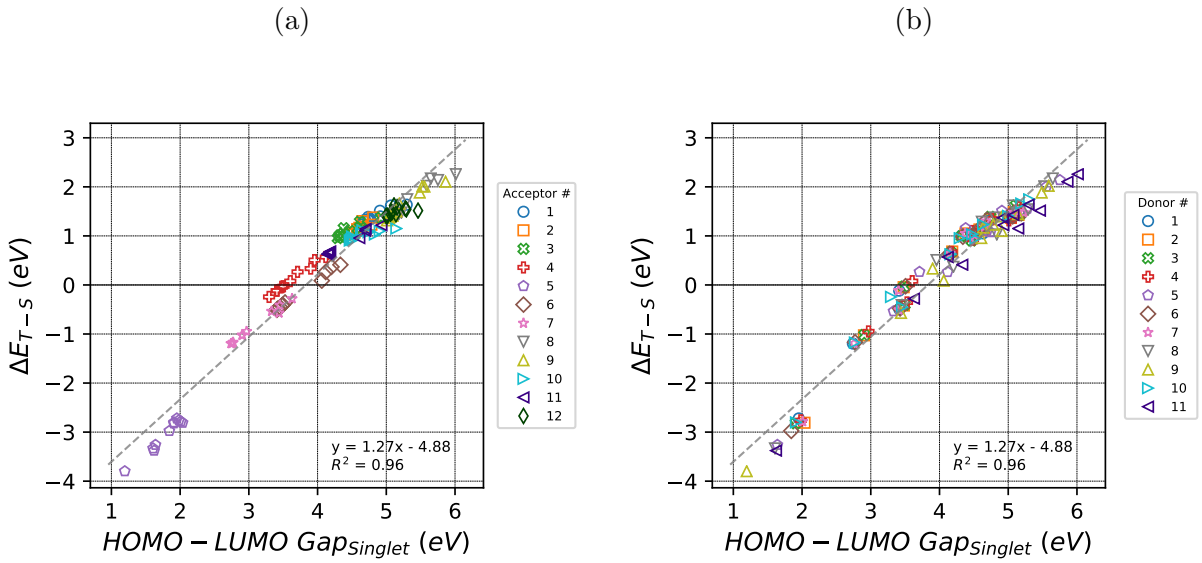


Figure 3.3: Correlation plots between the difference of the Triplet and Singlet energies of each oligomer versus its the HOMO-LUMO gap of the singlet species, both in eV, grouped by **a** the acceptor number and **b** the donor number. Linear best-fit line is shown as a dashed gray line.

the triplet state becomes increasingly stabilized. In some systems, broken-symmetry DFT calculations can predict the energetics of the spin-polarized singlet, although they require specifying a particular atomic radical center.[106]. In contrast, these oligomers access the triplet state specifically because of the delocalized π system. Furthermore, multi-reference methods such as MR-MP2[107]) would not properly capture the small HOMO-LUMO gap which stabilize the ground state triplet state. Consequently, we believe these systems will be useful target compounds for continuing electronic structure method development.

For comparison, and to further strengthen our conclusions, we also ran single-point energy calculations using the CAM-B3LYP functional[108] (Figure B1). While the scale of the axes have slightly changed, as well as the crossing point between singlet and triplet stability, a similar linear relationship between the HOMO-LUMO gap of the singlet species and ΔE_{T-S}

is observed.

From this correlation we can see that acceptor A5 appears to have the lowest singlet HOMO-LUMO gap and the most stable triplet state, followed by acceptor A7. Acceptors A4 and A6 both have several oligomers with stable triplet ground states; six oligomers for acceptor A4 and eight oligomers for acceptor A6. Interestingly, acceptor A5, which is the hydrogenated version of acceptor A4, was possibly made due to a human error. Nonetheless, this unintentional discovery resulted in a monomer that promoted a stable ground state triplet. However, we acknowledge that, realistically, the synthesis of acceptor A5 might prove to be a difficult endeavor.

In addition, we have examined the correlation between the singlet HOMO-LUMO gap calculated using the ωB97X-D DFT method and the GFN2-xTB semi-empirical method (Figure B2). Oligomers that contain acceptor A5 are outliers, showing no correlation with the rest of the set. Due to the low synthetic viability of acceptor A5, we removed those oligomers from the comparison. We fit the data points to linear, logarithmic, and radical functions, and found that the logarithmic function has the highest R^2 at 0.96, compared to 0.89 and 0.94 for the linear and square root functions, respectively. In short, while the HOMO-LUMO eigenvalues from density functional theory are unphysical, and gaps from ωB97X-D may be too large, and from GFN2-xTB may be too small compared to experiment, there is still a strong correlation between the two computational methods. This correlation can thus be used in the discovery of new ground state triplet oligomers, as only a single semi-empirical calculation is needed, skipping multiple time-consuming DFT calculations.

3.3.3 Monomers HOMO-LUMO

To further consider why the stability of a triplet ground state is dependent on the acceptor identity, the HOMO and LUMO eigenvalues of each acceptor and donor monomers were calculated, following the same process as the oligomers, as described in the **Methods** section (Figure B3). The HOMO eigenvalues on the donor monomers are, of course, generally less negative (above -8 eV) than the HOMO energies of the acceptor monomers (below -8 eV, with the exception of acceptors A5, A10, and A12.) Acceptors A4, A6, and A7, which

yielded some oligomers with ground-state triplets, have similar HOMO energies at -8.30 eV, -8.34 eV, and -8.34 eV respectively. Acceptor A11 has a similar HOMO eigenvalue, at -8.32 eV, but does not show triplet ground-state stability.

Acceptor A5 is the exception in this case, as it has a relatively high HOMO eigenvalue, at -6.66 eV — higher than all the other acceptor and donor monomers. This suggests that acceptor A5 might act as a very good donor. It would be interesting, to find other synthetically-accessible monomers with similar HOMO and LUMO eigenvalues and compare their performance to acceptor A5. However, this is beyond the scope of this work.

3.3.4 Strategies to Lower the HOMO-LUMO Gap

As we have shown, the most reliable predictor for the stability of the triplet ground state is the HOMO-LUMO gap of the singlet species. We therefore considered different general strategies to lower gap. Previous studies have shown that the identity of the heteroatom in the backbone of an oligomer as well as the identity of the side group affect the HOMO-LUMO energy gap[109, 110]. We chose sulfur (S), nitrogen (N) and selenium (Se) as the representative heteroatom of a 5-membered aromatic heterocycle; since we expected the HOMO-LUMO gap of the oligomer to decrease as the HOMO-LUMO gap of the heteroatom decreases[111, 109]. The side group, always on the 3- and 4- position in the heteroycle, representatives are a fused benzene ring for its potentially stabilizing effect on the quinoidal form[112], an ethylenedioxy group for its electron-donating effects[113], and no side group as a reference.

This set consists of 12 monomers — pyrrole, thiophene, selenophene, their 3-4-ethylenedioxy-derivatives (EDOP, EDOT, and EDOS respectively) and their benzo- derivatives (BP, BT, and BS respectively). In addition, we constructed a quinoidal version of the thiophene-based oligomers by forcing inter-monomer double bonds by adding methylidene ($H_2C=$) terminating groups (q-Thiophene, q-EDOT, and q-BT respectively), as can be seen in Figure 3.4a. By forcing a quinoidal bonding structure we can have a straightforward comparison with the aromatic bonding structure. We constructed the hexamer of each system, as previous studies show a high correlation with the calculated electronic energies of longer oligomers[55, 114],

and followed the same geometry optimizations and single-point calculations of both singlet and triplet species as described in the **Computational Methods** section below.

Figure 3.4b, demonstrates that lower singlet HOMO-LUMO gap correlates with a lower ΔE_{T-S} and a more stable triplet ground state. As expected, we also see correlation between the singlet HOMO-LUMO gap energy of the heteroatom and the singlet HOMO-LUMO gap energy of the hexamer, where nitrogen yields a higher gap, followed by sulfur, and lastly selenium with the lowest gap, consistent with previous results.

However, the side group identity has a large effect on the HOMO-LUMO gap energy as well. Figure 3.4b illustrates how the benzo- derivatives have a significantly lower HOMO-LUMO gap energy than the ethylenedioxy- derivatives, which itself has only a moderate reduction of the HOMO-LUMO gap energy over the parent oligomers with no side group.

In contrast to the findings above, the forced-quinoidal form of the thiophene-based oligomers show an opposite trend compared to its aromatic counterpart. While the quinoidal form of thiophene hexamer has a significantly lower HOMO-LUMO gap than the aromatic thiophene hexamer and a more moderate reduction compared to the ethylenedioxy- derivatives, this relationship reverses when it comes to the benzo- derivatives. While some promote a quinoidal bonding scheme as a strategy to induce a triplet ground state[81, 80, 96], we show here that this is not the case. This conclusion reinforces the discussion above, that a pro-quinoidal molecular design is guaranteed to make a polymer with a stable tripled ground state. The best overall strategy is to use monomers that promote a lower HOMO-LUMO gap, either by conjugation or inductively.

3.4 Conclusions

In conclusion, polymers with a triplet ground state have unusual and interesting properties that can open the door to novel and exciting applications. In this study we considered various strategies to produce such materials. We examined the correlation between the calculated stability of oligomers with a triplet ground state versus various electronic and geometric properties. We have found a high correlation between the energy difference between

(a)

(b)

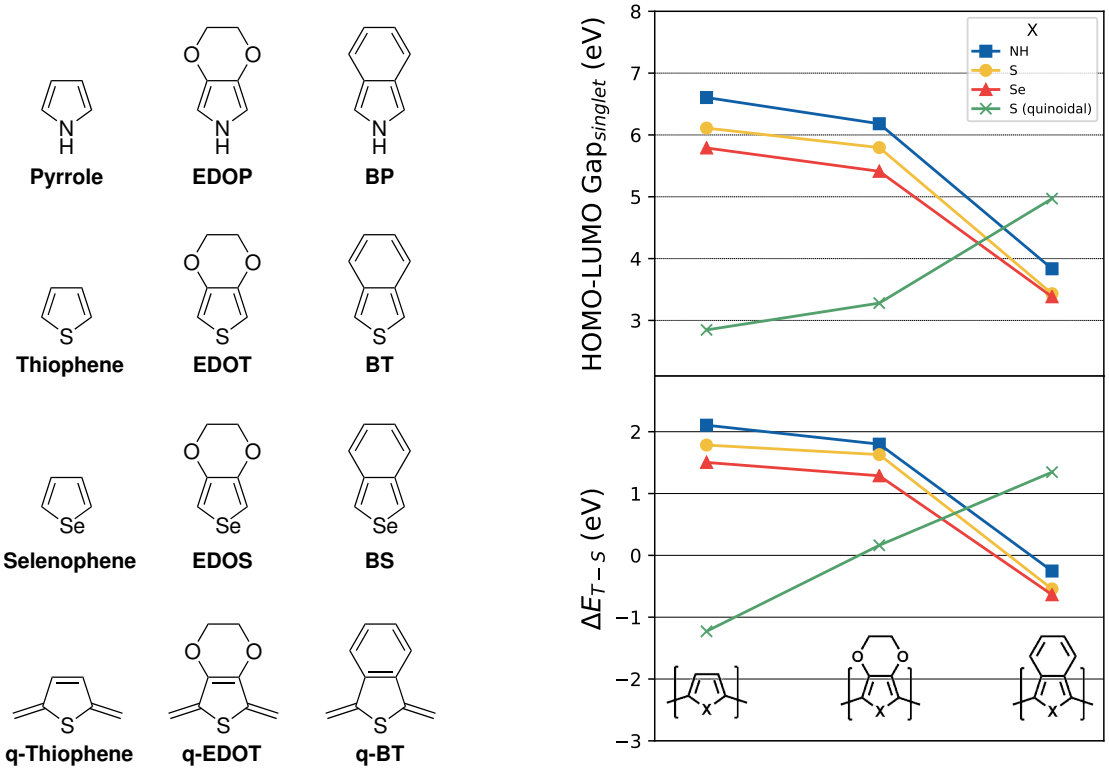


Figure 3.4: **a** The 12 monomers used in finding a strategy to lower the HOMO-LUMO gap — Pyrrol, 3-4-Ethyldioxypyrrole (EDOP), Benzopyrrole (BP), Thiophene, 3-4-Ethyldioxythiophene (EDOT), Benzothiophene (BT), Selenophene, 3-4-Ethyldioxyselenophene (EDOS), Benzoselenophene (BS), and the quinoidal versions of the thiophene-based monomers – denoted with "q-", **b** The singlet HOMO-LUMO gap of the hexamers (on the top), and their ΔE_{T-S} (on the bottom), both in eV. The different monomers, with "X" denote the different heteroatom as shown in the legend, are on the x-axis.

the triplet and singlet states of the oligomer (ΔE_{T-S}) and the HOMO-LUMO gap of the singlet state. We can use this correlation in order to find new candidates with triplet character by calculating just the HOMO-LUMO gap of the singlet. Moreover, general strategies to produce low band gap π-conjugated polymers can be used to find novel ground-state triplet materials.

Again, spin-polarized singlet states may complicate the correlation, a strong trend is observed between the decreasing HOMO-LUMO gap with two functionals, and the increasing stability of the triplet ground state. Additional development of electronic structure methods to treat these delocalized compounds and particularly the spin-polarized singlet state would be desirable.

While pro-quinoidal design strategies may yield ground-state triplet polymers, our examination over 132 oligomers suggests there is no overall correlation between quinoidal bond character and the stability of the triplet state. In addition, we have found that the HOMO and LUMO energies of the donors and acceptor monomers are also poor predictors for the stability of the triplet ground state. Instead, heteroatom substitution (e.g., sulfur to selenium) and side-group substitutions appear to yield increased stability of the triplet state without necessarily inducing quinoidal character.

These design rules can, in turn, be used for both experimental and computational design of new ground-state triplet conjugated materials. For example, use of a genetic algorithm or other generative method can sample a large number of potential oligomers through computational design.[84, 61]

3.5 Computational Methods

As mentioned above, the data set consists of 11 donor monomers and 12 acceptor monomers (Figure 3.1) studied by the Azoulay group[97, 98, 99, 100, 101, 92]. Long alkyl chains were replaced with shorter methyl groups in order to reduce computational needs with negligible effect on the electronic properties. A tetramer (octamer in Azoulay's notation) of every possible donor-acceptor pair (i.e. DADADADA, where D = donor monomer and A =

acceptor monomer) was created from their respective Simplified molecular-input line-entry system (SMILES) string[115, 116, 117], giving a set of 132 oligomers.

The geometry optimization of every oligomer was done in steps in order to reduce computational costs, starting with a conformer search and optimization using MMFF94[118] or UFF[119] with OpenBabel version 3.1.0[61], followed by GFN2-xTB[54], and ending with the B97-3c DFT functional[120] using Orca version 4.2.0[121, 122].

The single-point energy of each oligomer was calculated on the final optimized B97-3c geometry using the dispersion-corrected ωB97X-D3 DFT functional[123] with the def2-SVP basis set[124] using Orca. This process was done for both singlet and triplet species of each oligomer. Single-point energy calculations using the dispersion-corrected CAM-B3LYP functional were done in the same way. The output files were processed using Python with the cclib library[125] for the electronic, highest occupied molecular orbital (HOMO) and lowest unoccupied molecular orbital (LUMO) eigenvalues, and the Open Babel library[61] for the extraction of the inter-monomer bond lengths. The difference between the triplet and singlet energies was calculated as

$$\Delta E_{T-S} = E_{Triplet} - E_{Singlet} \tag{3}$$

4.0 Using Genetic Algorithms to Discover Novel Ground-State Triplet Conjugated Polymers

This chapter is adapted from:

Omri D. Abarbanel, Geoffrey R. Hutchison; Using genetic algorithms to discover novel ground-state triplet conjugated polymers. *Physical Chemistry Chemical Physics* 2023, 25, 11278-11285. DOI: doi.org/10.1039/D3CP00185G.

It is a collaborative effort in which the author implemented the algorithm, performed the calculations and data analysis, generated the figures, and wrote the manuscript; G.R.H. conceived and directed the project.

4.1 Summary

Stable ground-state triplet π-conjugated copolymers have many interesting electronic and optoelectronic properties. However, the large number of potential monomer combinations makes it impractical to synthesize or even just use density functional theory (DFT) to calculate their triplet ground-state stability. Here, we present a genetic algorithm implementation that uses the semi-empirical GFN2-xTB to find ground-state triplet polymer candidates. We find more than 1400 polymer candidates with a triplet ground-state stability of up to 4 eV versus the singlet. Additionally, we explore the properties of the monomers of those candidates in order to understand the design rules which promote the formation of a stable ground-state triplet in π-conjugated polymers.

4.2 Introduction

Although organic π-conjugated polymers have been researched for their unique electronic properties and potential uses[17, 16], a new subclass of π-conjugated organic polymers with

a triplet ground-state has recently been introduced. Research into the discovery of these ground state triplet organic π-conjugated polymers has increased in the last few years, with works on molecular design and characterization. While many of these efforts are still ongoing, they all follow similar design rules, copolymers composed of electron accepting and electron donating monomer pairs[92, 95, 96, 80, 81].

By choosing the right combination of monomers a polymer with a small energy gap between the highest occupied molecular orbital (HOMO) and the lowest unoccupied molecular orbital (LUMO), i.e. the HOMO-LUMO gap, can be achieved. This leads to degenerate or near-degenerate frontier molecular orbitals and allows for the unpairing of the normally paired HOMO atoms into two singly-occupied molecular orbital (SOMO) with a biradical nature[81, 91].

Previous computational work has shown that the size of the HOMO-LUMO gap of singlet species directly correlates with the stability of the triplet species. That is, polymers with a small HOMO-LUMO gap also have a more stable triplet ground state. Additionally, the identity of the acceptor monomer was found to have a high correlation with the stability of the triplet ground state, where polymers that share the same acceptor monomer will have similar electronic properties[126].

Experimentally, it is highly impractical to create and characterize every monomer combination. Even computationally, this can require many resources in order to comb through a large number of monomer pairings. This calls for a more efficient method that can sort and eliminate unwanted monomers while preserving those that show promising results.

There are many approaches that can be used to accelerate the discovery of new materials. Machine learning (ML) is one method that is becoming increasingly popular in this field[127, 128, 129]. However, it requires a large data set, which does not exist in this case. Another method is the genetic algorithm[56] (GA) which is a non-deterministic optimization algorithm. The GA is an iterative method, where each generation a new set of offspring are created from the previous generation surviving parents, and those who survive the fit test go on to be the parents in the following generation. The algorithm also includes the possibility of a random mutation that can result in an increase or decrease in the survival rate.

In this work, we used a genetic algorithm to discover new ground-state triplet polymers.

Our data set consists of 1226 monomers, which, if thoroughly combined with each other, would create over 1.5 million (1226^2) potential polymers to work with. This comprehensive method can be associated with high computational resources and time. The iterative GA method can efficiently sift through the large number of combinations and produce stable ground-state triplet candidates with high confidence. Moreover, multiple GAs can run in parallel, increasing the verity of potential candidates.

Additionally, from the GA we can also produce a list of the most common monomers that were used in each GA run. A highly common monomer means that it survived natural selection and passed to the next generation. Also, by promoting a low HOMO-LUMO gap, a monomer has a higher chance of creating offsprings. From these common monomers we can gain insights into which monomer properties correlate with a stable triplet ground-state in the full polymer.

4.3 Methods

4.3.1 Correlation Between the Singlet HOMO-LUMO Gap and Stability of the Triplet State

In a previous study, we found a strong linear correlation between the HOMO-LUMO gap of the oligomer singlet state and the stability of its triplet[126]. That is, oligomers with a low HOMO-LUMO gap lead to a stable open-shell electronic structure due to the frontier molecular orbitals being closer energetically. Oligomers with a triplet ground-state tend to have a biradical electronic structure, in which each electron is found in a separate singly-occupied molecular orbital (SOMO).

Using these findings halves the number of potential calculations that needs to be performed to find if an oligomer has an open-shell electronic structure, as only the HOMO-LUMO gap of the singlet species is needed. This completely negates the necessity of calculating the electronic energy of the triplet state, and drastically accelerates the discovery of open-shell π-conjugated materials.

Furthermore, the oligomers in the aforementioned study were constructed from pairs of electron donor and electron acceptor monomers. However, we discovered that the stability of the triplet ground-state is based primarly on the identity of the acceptor monomer. By expanding the list of electron-accepting monomers, we can find better oligomers with a biradical nature.

4.3.2 Correlation between GFN2-xTB HOMO-LUMO gap and ωB97X HOMO-LUMO gap

In addition to the correlation between the singlet HOMO-LUMO gap and the stability of the triplet state, we also previously found a correlation between the singlet HOMO-LUMO gap calculated using density functional theory (DFT) and the HOMO-LUMO gap calculated using the semi-empirical GFN2-xTB method. While the correlation is imperfect, there is a clear trend. To strengthen the correlation, we calculated the HOMO-LUMO gap using both GFN2-xTB and ωB97X-D3 of randomly generated list of new oligomers from the expanded list of monomers used in this study (Figure C1). See Section 4.3.3.2 for how these calculations were performed.

Although both the logarithmic and radical functions can describe the correlation, the radical function slightly better ($R^2 = 0.83$) than the logarithmic ($R^2 = 0.74$), both show that a small HOMO-LUMO gap calculated with GFN2-xTB would correlate with a small HOMO-LUMO gap calculated using DFT. As GFN2-xTB is a semi-empirical method, it is much faster than DFT and can greatly accelerate the GA. We therefore use GFN2-xTB–calculated HOMO-LUMO gap in the GA, and find oligomers that minimize the HOMO-LUMO gap with every generation.

4.3.3 Computational Methods

4.3.3.1 The Genetic Algorithm

Genetic algorithms follows Charles Darwin's *Survival of the Fittest* idea, which describes how evolution works[130]. Parents with certain traits create offspring that end up with some

combination of those traits. Offsprings with combinations of traits that help them survive in their environment can produce their own offsprings. This cycle can continue indefinitely or terminate by some external force. At certain points during this process a new, never seen before, trait has a chance to appear due to a random mutation. These mutations can either have no effect, help, or hinder the survival and reproduction chances of a off-spring.

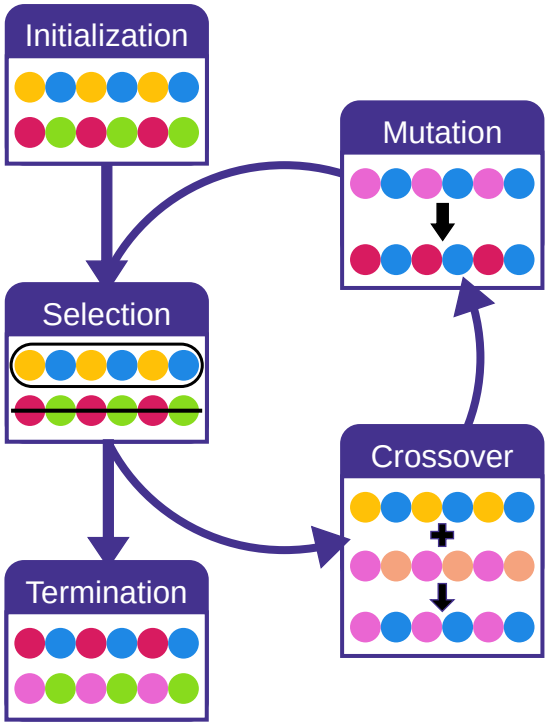


Figure 4.1: The five steps of the genetic algorithm - initialization, selection, crossover, mutation, and termination. The selection-crossover-mutation cycle is repeated a set number of times before stopping at the termination step.

Our genetic algorithm follows several steps (Figure 4.1):

1. Initialization — which creates a starting population for the algorithm to work with. These are the "parents" in the GA.
2. Selection — which selects some of the population to survive and continue to reproduce, while it eliminates the others, based on some fitness function. This is the survival rate of the population.

3. Crossover — which creates a new population from a random combination of two parents from the population that survived the previous step. Here, successful parents create new candidates with a combination of their traits.
4. Mutation — in which, given some mutation rate, some of the current population has its traits changed to a randomly chosen one.
5. Termination — After repeating steps 2-4 for some number of cycles, the GA ends with the last surviving population.

In this study, the GA was initialized with a population size of 32 oligomers. Each oligomer was composed of a pair of monomers, repeated four times in an alternate fashion (i.e. ABABABAB, where A and B are the first and second monomers in the oligomer, respectively). During the *Selection* step, we use the GFN2-xTB–calculated HOMO-LUMO gap as the fitness function and eliminate half of the population, that is, 16 oligomers, with the largest HOMO-LUMO gap. In the *Crossover* step we create 16 new off-springs by randomly choosing two monomers from the surviving population, which brings the total population size back to 32. During the *Mutation* step, every oligomer has a 40% chance to have one of its monomers replaced by a random monomer from the entire list of possible monomers. We have used the same hyperparameters for the GA from a previous study done in our group as they have shown to be effective for similar molecular systems[131, 58]. The GA terminates the *Selection–Crossover–Mutation* cycle after 40 generations, which we have found to be sufficient in finding the minimal calculated HOMO-LUMO gap (Figure 4.2). However, since random chance is an integral part of the GA, a single run of the GA can miss many potential oligomers with a low HOMO-LUMO gap—unless the GA is left to run indefinitely, which is an impossible task. To save run time and increase the chance of finding oligomers with a low HOMO-LUMO gap, we ran the GA ten times in parallel.

4.3.3.2 Geometry Optimization and Single Point Calculations

The GA was implemented using the Python programming language, version 3.8[132] using custom code which can be found on GitHub at https://github.com/hutchisonlab/oligomer-ga. We ran the GA on a list of 1226 different organic monomers[131, 55, 133]. The full list

of monomers can be found in Appendix C. Each monomer is numbered from 0 to 1225 in no particular order. The oligomers were 8 monomers long, that is, 4 monomer pairs (that is, AB-AB-AB-AB, where A and B are the different monomers in the oligomer), in order to match previous studies[126, 81]. Additionally, this length was chosen because of the balance of a good approximation of the HOMO-LUMO gap of the long-chain polymer and the computational costs and time of running DFT calculations, which can take anywhere between a few days and a few weeks. Exploring the monomer sequence, i.e., in an alternating form, is beyond the scope of this study and can be the subject of future research. However, previous work in our group have studied the effect of the monomers sequence on the electronic properties of the oligomer, and can significantly tune HOMO-LUMO energetics[134, 135, 136, 137, 131].

In every step of the GA the oligomers were constructed from the SMILES strings of their respective monomers and followed by an initial force-field geometry optimization, and conformer search step with UFF[119] or MMFF94[118] using OpenBabel[61] version 3.1. A second geometry optimization step and the calculation of the HOMO-LUMO gap were done using GFN2-xTB[54] version 6.4.1. The GFN2-xTB output was parsed using a custom Python script.

The potential oligomers that were found by the GA were further analyzed underwent a third geometry optimization using the Density Functional Theory (DFT) B97-3c functional[120] followed by a single point calculation with the ωB97X-D3 functional[123, 138] and the def2-SVP basis set[124] using ORCA version 4.2.0[121, 122]. This process was repeated separately for both the singlet and triplet species of each oligomer. Single-point energy calculations using the dispersion-corrected CAM-B3LYP functional[139] were done in the same way. The energies and HOMO-LUMO gaps calculated by Orca were parsed using the cclib Python package[125].

The electronic energy difference between the singlet and triplet species of each oligomer or monomer is defined as

$$\Delta E_{T-S} = E_{Triplet} - E_{Singlet} \tag{4}$$

with $E_{Singlet}$ and $E_{Triplet}$ are the electronic energies of the singlet and triplet species, respectively. That is, when the ΔE_{T-S} if a certain oligomer is negative, its triplet ground-state is

more stable, and vice versa.

4.4 Results and Discussion

4.4.1 The Genetic Algorithm

As mentioned before, we ran the GA ten times in order to diversify the potential list of monomers with a small HOMO-LUMO gap. The objective of our GA, that is, the value it was aiming to optimize, was a small HOMO-LUMO gap. As seen in Figure 4.2 the GA did do as intended and indeed minimized the GFN2-xTB–calculated HOMO-LUMO gap. The average HOMO-LUMO gap (Figure 4.2 Top) shows a significant decrease from the first few generations, while later it is subject to some randomness due to the nature of the GA. However the lowest HOMO-LUMO gap (Figure 4.2 Bottom) shows a clear trend where the GA does finds oligomers with a very small HOMO-LUMO gap.

From the bottom figure in Figure 4.2 it can be seen that after about 40 generations all the runs converged on a very low HOMO-LUMO gap, within the limitations of the GA. Although each run started with a random set of oligomers with different HOMO-LUMO gaps, they all ended with a set of oligomers that, on average, have a lower HOMO-LUMO gap and at least one oligomer with a GFN2-xTB–calculated HOMO-LUMO gap less than 0.01 eV. Based on our observed correlation between the GFN2-xTB and ωB97X-D3 gaps, these are roughly equal to a gap of 1.5 eV. 1426 copolymer candidates have been found to have a GFN2-xTB–calculated gap of less than 0.1 eV, which correspond roughly to a ωB97X-D3 gap of 2.8 eV.

4.4.2 Top Oligomers

The top 20 oligomers, that is, the oligomers with the smallest GFN2-xTB–calculated HOMO-LUMO gap found in any of the GA runs, were extracted for further analysis. The HOMO-LUMO singlet gap and the electronic energies of the singlet and triplet states were calculated using the ωB97X-D3 functional following the steps described in Section 4.3.3.2.

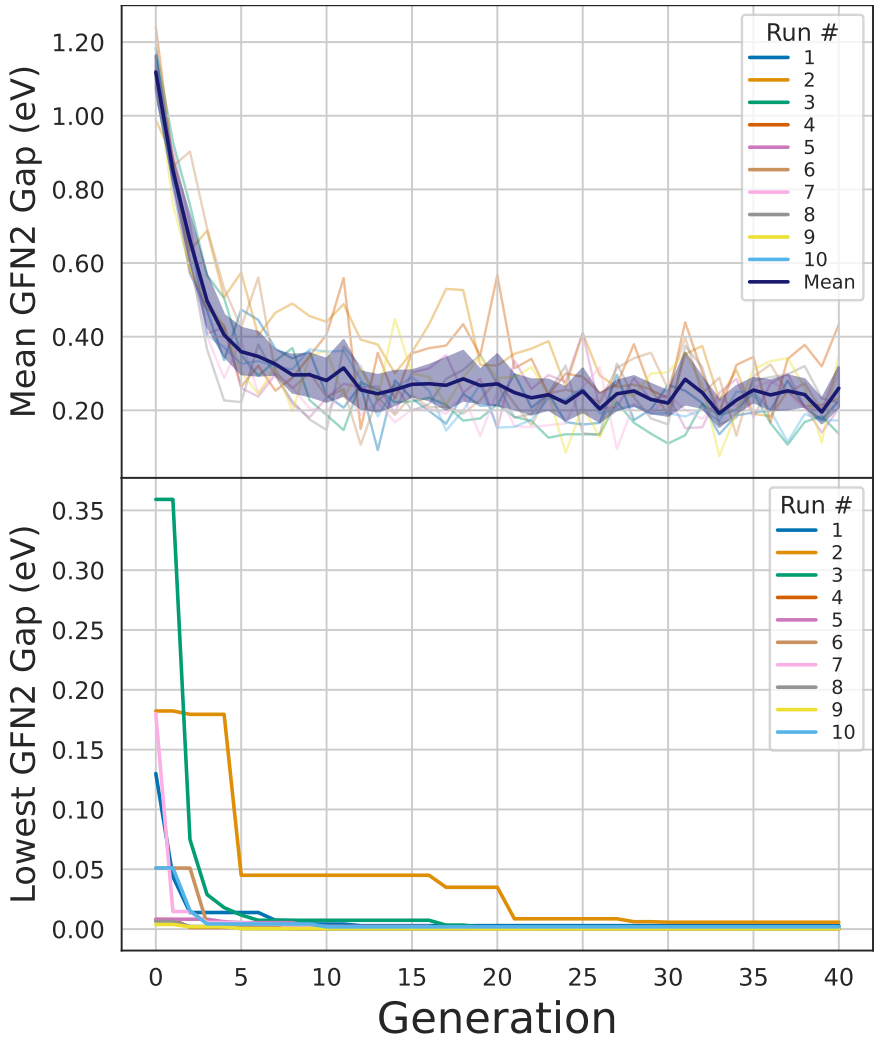


Figure 4.2: Top: the mean GFN2-xTB–calculated HOMO-LUMO gap in each generation for each GA run, with the mean gap and standard deviation for each generation over all runs in dark blue. Bottom: the lowest GFN2-xTB–calculated HOMO-LUMO gap of every generation in each GA run.

Of the top 20, four oligomers encountered various problems during one or more of the DFT calculation steps and, therefore, were removed from further analysis. All of the top 20 oligomers share one of these two monomers, 35 or 642 (Figure C2). Both monomers share a similar molecular structure, as can be seen in Figure 4.3 Bottom, except that monomer 642 includes a vinyl bridge.

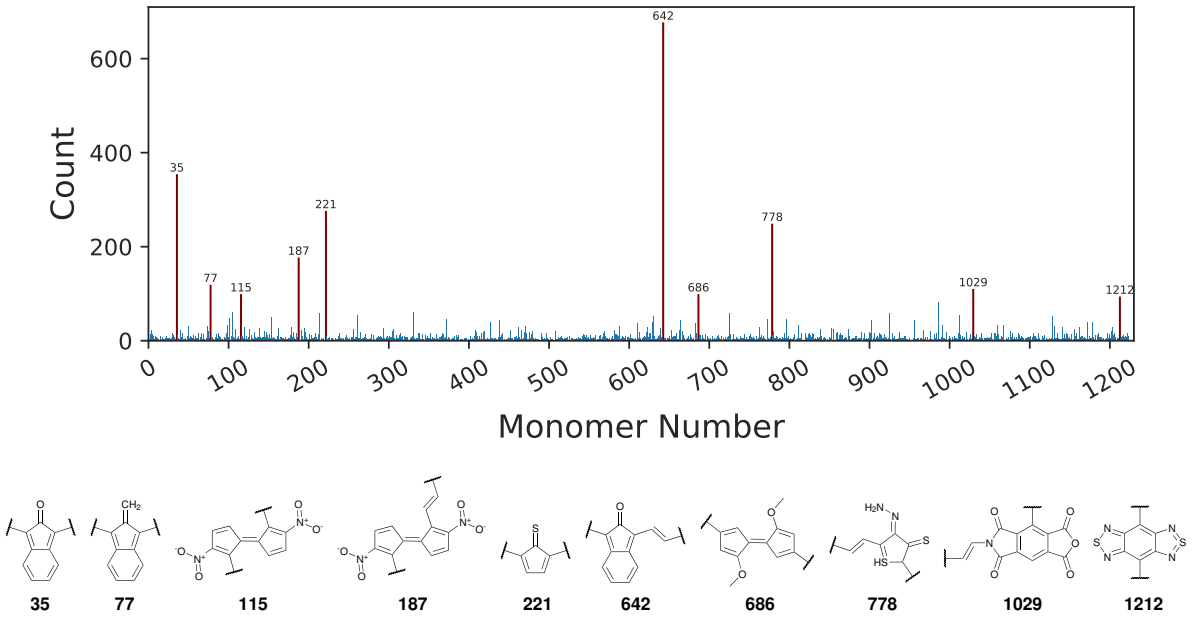


Figure 4.3: Top: The number of times a monomer has been used in any of the ten GA runs. The top 10 most common monomers are boldly emphasized in red and have their monomer number above them. Bottom: The structures of the top 10 most common monomers.

In Figure C3 it can be seen that the ωB97X-D3–calculated HOMO-LUMO gap and electronic energies agree with the GA and the GFN2-xTB–calculated values, as those oligomers indeed have a small singlet HOMO-LUMO gap and a stable triplet ground-state. This further shows that GFN2-xTB can be a good surrogate for DFT in finding oligomers with small HOMO-LUMO gaps with the GA. While most of the oligomers show similar electronic properties, two of them, one constructed from monomers 642 and 365 and another constructed from monomers 642 and 128, look like outliers. However, we expected to see some variation between the singlet HOMO-LUMO gap and the ΔE_{T-S} since the correlation is not perfect. On the contrary, those two outliers are the two data points closest to the best-fit

line calculated in the previous study[126].

Moreover, the outlier 128-642 can be attributed to its conformation, since its lowest energy conformation found during the geometry optimization steps had the oligomer folded on itself instead of the flat linear conformation the other oligomers showed (Figure C4a). We modified the conformation by manipulating the bond angles to create a more linear conformation using Avogadro2 version 1.95.1, followed by the same geometry optimization and single point calculations as described in section 4.3.3.2, The resulting geometry remained in the modified flat conformation (Figure C4b), and the HOMO-LUMO gap of the oligomer decreased from 1.04 eV to 0.35 eV and its ΔE_{T-S} also decreased from -2.82 eV to -3.23 eV, and it got closer to the cluster of the other oligomers (Figure C3). The outlier 365-630 had the same flat and linear conformation as the other oligomers; however, its backbone includes a 7-membered ring, which breaks aromaticity and disrupts conjugation. We hypothesize that this may be the cause of this oligomer's properties. Nonetheless, we wish to reiterate that those two outliers still exhibit a stable triplet ground-state.

We would like to add that while we expect the conformation of the oligomers to be extended due to the rigidity that comes from the conjugated π-system , there is a possibility that the lowest energy conformation of an oligomer would be a nonlinear one, as seen above for oligomer 128-642. The conformation of the oligomer does affect the HOMO-LUMO gap, and we see that the HOMO-LUMO gap decreased when the oligomer conformation was intentionally extended. However, the HOMO-LUMO gap dispersion is relatively small compared to the scale of the triplet ground state stabilization [140, 141, 142].

In principal, quantum calculations for HOMO-LUMO gap and singlet-triplet energies should use a substantial conformational search, followed by a Boltzmann-weighted average of properties. In practice, given the size of the conjugated systems included, proper conformer sampling (e.g., with GFN2)[53] would significantly increase the run-time of the GA.

Additionally, the spin density plots of the top 20 oligomers (Figures 4.4 for an example and C6 for the rest of the oligomers) show their biradical nature by the delocalization of the two unpaired electrons. It can be seen, in some oligomers easier than others, that the oligomers show higher spin density towards the edges of the triplet ground-state oligomer. This matches with previous computational studies that showed a similar effect on a poly-

mer that has been experimentally synthesized and characterized as having a triplet ground-state[81]. This effect occurs because of the Coulomb repulsion forces as the two unpaired electrons have the same spin in the triplet state.

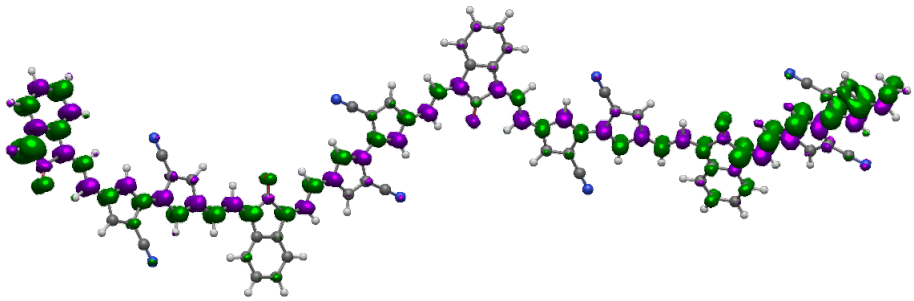


Figure 4.4: Spin density plot of the oligomer constructed from monomers number 642 and 630. The purple and green orbital colors correspond to the α and β electrons, respectively. Isosurface value is 0.002 a.u.

[143]

To see if there are design rules that can be used to find other polymers with a triplet ground state, we have extracted the top oligomers with a GFN2-xTB–calculated HOMO-LUMO gap of less than 0.2 eV that were generated in the 10 GA runs (2024 oligomers), and compared them to all of the possible ($\sim$ 1.5 million) oligomer in our dataset. To help generalize the design rules we used RDKit to calculate various descriptors on the monomer pairs, instead of the full length oligomer . The descriptors include the molecular weight of the pair, the number of atoms, number of rotatable bonds, number of hydrogen-bond acceptors and donors, number of rings and aromatic rings, the partition coefficient (Crippen $logP$[144]), π-system size[55], and number of nitrogen, oxygen, sulfur, selenium, and halogen atoms in the pair (Figures C11 and C12). Although there is some noise due to the relatively small sample size in Figure C12, it is still possible to extract some potential design rules. For example, there are more oxygen atoms in the top monomer pairs (2.7 ± 2.1) compared to the full monomer combinations (1.7 ± 1.6), as well as a high proportion of monomer pairs with a π-system size of 12 atoms, while other comparisons can be attributed to the small sample size. However, this analysis shows that there are no generalized design rules, and that a search algorithm, like this GA, is needed in order to traverse the vase chemical space

and find potential polymers with a triplet ground state.

4.4.3 Top Monomers

From the analysis of the top oligomers from the GA, two monomers have been shown to be ubiquitous, 35 and 642 (Figure C2). This further demonstrates that the stability of the triplet ground-state is frequently controlled by one of the two monomers in the oligomer. That is, some monomers induce a small HOMO-LUMO gap in many of the oligomers they are found in when paired with many different monomers, and we expect these monomers to be more common in the GA. To find which monomers in the set exhibit similar properties, a histogram of the number of occurrences of each monomer in all GA runs was constructed (Figure 4.3 Top). A higher number of instances in the GA would suggest a higher survivable rate throughout the GA cycle, due to it contributing to a small HOMO-LUMO gap relative to the rest of the population.

In fact, it appears that of the 1226 monomers in our data set, only a small subset has been captured by the GA to promote a small HOMO-LUMO gap. In Figure 4.3 the top 10 most common monomers are highlighted, along with their molecular structures at the bottom. At a first glance some of those monomers show a common electron-accepting motifs, such as monomers 115 and 187 with two highly electron withdrawing nitro groups, or monomer 1212 which is another common acceptor monomer used in various π-conjugated polymers[81, 126].

Another common motif in the top 10 monomers is the vinyl bridge, also called a vinylene link. The inclusion of a vinyl bridge in the polymer backbone has been shown to lower the HOMO-LUMO gap by extending the conjugation of the π-system, leading to greater delocalization of πelectrons[145]. As a testament to this hypothesis, our GA found monomers with a vinyl bridge (e.g., monomers number 642 and 187) at a higher frequency than their derivatives without a vinyl bridge (monomers 35 and 115, respectively), as shown in Figure 4.3. Moreover, Cordaro and Wong have also commented that in their experience, in addition to drastically decreasing the HOMO-LUMO gap, a vinyl bridge also improves the solubility of polymers due to the increase in the polymer flexibility[145]. Therefore, polymers with a low HOMO-LUMO gap, and potentially a stable triplet ground-state, will benefit from

including a vinyl bridge in their backbone—both by lowering the intrinsic HOMO-LUMO gap compared to the non-bridged version and by potentially improving polymer solubility for synthesis, characterization, and application.

4.4.4 Monomer Properties

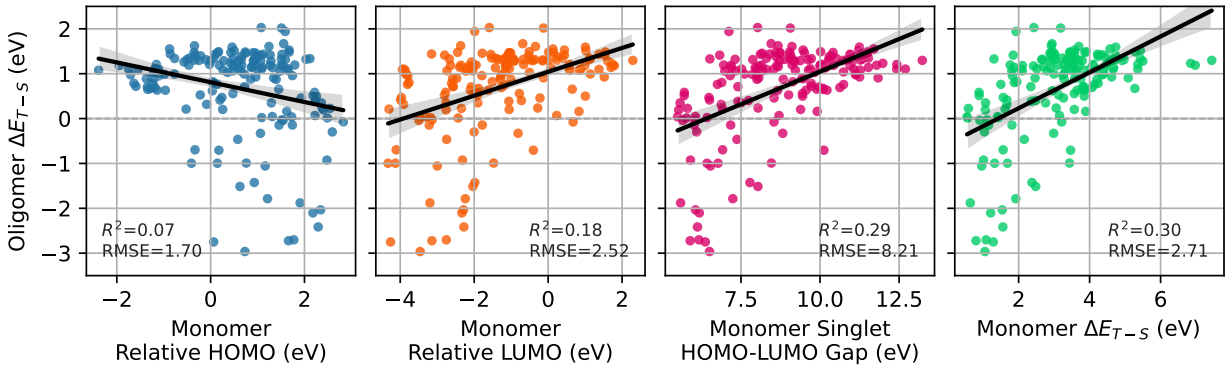


Figure 4.5: Monomers' HOMO level (relative to thiophene's), LUMO level (relative to thiophene's), HOMO-LUMO gap, and their triplet ground-state stability (ΔE_{T-S}) versus the stability of the oligomer's triplet ground-state stability when paired with monomer 630. The linear best-fit line and standard deviation are shown in black line in each plot, as well as the R^2 and the RMSE (in eV).

From looking at the list of the top 10 monomers, the first questions that should be asked are what are the electronic properties of those monomers have in common and whether, by discovering this property, we can find other monomers that share it and promote a stable open-shell electronic structure since similar monomers should have similar properties.

As mentioned before, we previously showed that the identity of the acceptor monomer has the highest correlation to a stable triplet ground-state[126]. However, the absolute classification of monomers between acceptor and donor is vague because these are relative terms. These are usually described by their HOMO levels, as a donor monomer will have a high HOMO level, while an acceptor will have a small one. However, an absolute scale is difficult to derive, as a monomer with a low HOMO level compared to its oligomer counterpart

will behave as an acceptor (Figure C5). A higher difference between the HOMO levels would entail a stronger donor-acceptor pair and vice versa. f To classify the monomer into strong and weak acceptors and donors, we needed to set a relative scale because absolute HOMO eigenvalues highly depend on the DFT functional and the basis set used. Some have used thiophene as a *"spacer"* monomer in various π-conjugated polymers, as it is claimed to not affect the electronic properties significantly[146, 147, 148, 149]. For the same reason, some have used thiophene as a reference monomer when comparing different donor and acceptor monomers. Therefore, we examine the relative HOMO and LUMO eigenvalues, as well as the HOMO-LUMO gap and the stability of their triplet state (ΔE_{T-S}) for all monomers (Figure C7). The single-point calculations using the ωB97X-D3 functional followed the same steps as the full oligomers, as described in Section 4.3.3.2.

For comparison and to reaffirm our results, we also performed single-point calculations on all monomers using the dispersion-corrected CAM-B3LYP functional (Figure C8). The results show similar distributions compared to the ωB97X-D3 single point calculations (Figure C7), showing that those results appear to be consistent across multiple functionals.

The top ten most common monomers in Figure 4.3 have small HOMO-LUMO gaps and small ΔE_{T-S}, as well as relatively low LUMO levels while their relative HOMO levels are more spread out (Table C1). However, these monomers are not all in the extreme ends of any category, and there are other monomers with small HOMO-LUMO gaps, for example, that did not show up as common in the GA as those top ten monomers. We can attribute this to several potential causes:

- Due to random chance in the GA. The GA is a stochastic optimization method, and by chance some potentially good monomers were not selected.
- Due to the electronic structure of the monomer. Some monomers with a small HOMO-LUMO gap, for example, have an antiaromatic electronic structure—like monomers with fused alternating 5- and 6-membered rings, such as s-indacene. This, we hypothesize, inhibits conjugation in the oligomer and does not promote a small HOMO-LUMO gap.
- Due to a human error with the SMILES string of the monomer. Some of the SMILES strings might have the wrong polymerization site which can break aromaticity and conjugation when the monomer is part of an oligomer.

- Due to inaccuracies in the GFN2-xTB calculations. While we have found a correlation between GFN2-xTB and ωB97X-D3 HOMO-LUMO gaps (Figure C1), this correlation is not as strong for small HOMO-LUMO gaps. It is possible that due to this some potentially good monomers did not survive the *Selection* step in the GA. This is a trade-off that we accept to greatly accelerate the GA.

Similarly to the full-length oligomers, a high correlation ($R^2 = 0.86$) was found between the monomers' singlet HOMO-LUMO gap and the stability of the triplet ground-state (Figure C9). This agrees with previous studies that showed a biradical nature in molecules with a small HOMO-LUMO gap[104, 93, 94, 150, 151].

4.4.5 Other Potential Monomers

To find which property contributes the most to the stability of an open-shell electronic structure in the oligomer, as well as other monomers that the GA might have missed, we looked at four different monomer properties: relative HOMO level, relative LUMO level, HOMO-LUMO gap, and their triplet ground-state stability (ΔE_{T-S}). A representative selection of monomers with a range of values for each property were selected, and an oligomer was created for each monomer and monomer 630—which was paired with monomer number 642 in the oligomer with the second most stable triplet ground-state (Figure C10). We chose to use monomer 630 over monomer 365, which was paired with monomer 642 and had the most stable triplet ground state, since it contained a 7-membered ring in its backbone which broke its aromaticity and interrupted its conjugation to the πsystem. Monomer 630 has a highly conjugated and aromatic structure that includes a vinyl bridge, and we hypothesized that it will create more consistent and explainable results. The DFT singlet HOMO-LUMO gap and the singlet and triplet electronic energies for each oligomer were calculated as described in Section 4.3.3.2.

Figure 4.5 show the correlation between the monomers' property versus the oligomers' triplet ground-state stability. There is no strong correlation between each of the monomer properties and the stability of the whole oligomer, as they show heteroscedastic behavior. That is, monomers with low relative HOMO levels and high relative LUMO levels, HOMO-

LUMO gap, and ΔE_{T-S} show low triplet state stability in the oligomer. On the other end, monomers on the opposite side of those properties do not show a clear-cut correlation between the property and the oligomers ΔE_{T-S}, at least when paired with monomer 630.

Interestingly, while oligomers with monomer number 630 were not as ubiquitous in the GA as other monomers were, Figure 4.5 show that many oligomers that include monomer number 630 did show a stable triplet ground-state—including monomers that are not in the top 10 most common monomers in the GA (Figure 4.3). For example, the oligomer constructed from monomers 630 and 261 showed a very strong ($\Delta E_{T-S} = -2.96$ eV) triplet ground-state stability, while not being found in any of the GA runs. This very low ΔE_{T-S} would be comparable to the top 20 oligomers found in the GA (Figure C3). See Table C2 in the **Supporting Information** for the full data.

4.4.6 Some Remarks

The finding above highlights a weakness in genetic algorithms as a whole, due to their non-deterministic nature and stochastic behavior. In other words, GAs can find a local optima, while sometimes missing the global one. There are ways to mitigate this behavior by tuning the GA's hyperparameters, such as the population size, mutation rate, and elitism rate [58]. However, even with well-tuned hyperparameters, there is still a chance that the GA misses the global optima. While there are other, deterministic algorithms that can find the global optima, they come with a greater computational cost[152]. In our case we tried to avoid this problem by running the GA 10 times, but even so it is evident that the GA did miss some potential candidates. The likelihood of this happening can be reduced by running the GA for more generations and more times, but then the return-on-investment (ROI) might not be favorable if this takes longer and has higher computational costs.

Another point we want to emphasize here, as Figure 4.5 shows, is that the identity of one monomer does not correlate with the oligomer triplet ground state stability, and it is the combination of the two monomers that overall dictates the oligomer's properties. While we presented here various monomers that were common in the GA (Figure 4.3), not every oligomer with them had a small HOMO-LUMO gap. For example, monomer 642 was the

most common monomer in the GA and in the top 20 oligomers, but when combined with monomer 659 it had a GFN2-xTB HOMO-LUMO gap of 1.22 eV — which would correlate to ∼5.5 eV DFT HOMO-LUMO gap, according to Figure C1, and a more stable singlet ground state than a triplet one, according to Figure C3, by ∼2.0 eV.

4.5 Conclusions

Ground-state triplet polymers have a unique electronic structure and properties that have many possible uses in electronic devices. In this study we demonstrated how a Genetic Algorithm combined with GFN2-xTB, a fast semi-empirical method, can find unique and novel π-conjugated organic co-polymer candidates with a stable triplet ground-state. Those candidates exhibit a small HOMO-LUMO gap, which was previously shown to promote a triplet ground-state electronic structure due to the frontier molecular orbitals getting closer in energy. The spin densities show the biradical nature of those candidates, and the delocalization of the two unpaired electrons over two different singly-occupied molecular orbitals. DFT calculations show a triplet ground-state stabilization for up to 4 eV for the oligomer, and we expect this value to be similar or greater for the full-length polymer.

In addition, we have found that a small number of monomers have been found by the GA to promote a small HOMO-LUMO gap. All of those monomers exhibit small HOMO-LUMO gaps on their own, which helped promote a small HOMO-LUMO gap in the full oligomer. However, no other monomers with a small HOMO-LUMO gap were found by the GA, which shows that the GA has flaws. While the stochatstic nature of the GA imply that it can sometimes miss a potential candidate it is still a faster and more efficient method than an exhaustive search over the vast chemical space, particularly for finding top candidates and relevant motifs.

5.0 QupKake: Integrating Machine Learning and Quantum Chemistry for micro-pKa Predictions

This chapter is adapted from:

Omri D. Abarbanel, Geoffrey R. Hutchison; QupKake: Integrating Machine Learning and Quantum Chemistry for Micro-pKa Predictions. *The Journal of Chemical Theory and Computation* 2024. DOI: doi.org/10.1021/acs.jctc.4c00328.

It is a collaborative effort in which the author implemented the machine learning algorithm, performed the calculations and data analysis, generated the figures, wrote the Python package, and wrote the manuscript; G.R.H. conceived and directed the project.

5.1 Summary

Accurate prediction of micro-pK_a values is crucial for understanding and modulating the acidity and basicity of organic molecules, with applications in drug discovery, materials science, and environmental chemistry. This work introduces QupKake, a novel method that combines graph neural network (GNN) models with semiempirical quantum mechanical (QM) features to achieve exceptional accuracy and generalization in micro-pK_a prediction. QupKake outperforms state-of-the-art models on a variety of benchmark datasets, with root mean square errors (RMSEs) between 0.5-0.8 pK_a units on five external test sets. Feature importance analysis reveals the crucial role of QM features in both the reaction site enumeration and micro-pK_a prediction models. QupKake represents a significant advancement in micro-pK_a prediction, offering a powerful tool for various applications in chemistry and beyond.

5.2 Introduction

The acid-base dissociation constant (pK_a) is a fundamental physicochemical property of molecules, with broad applications in organic synthesis, environmental chemistry, medicinal chemistry, and drug design and discovery.[18, 19] The pK_a value of a molecule reflects its relative propensity to donate or accept a proton, and can have a significant impact on its solubility, membrane permeability, protein binding affinity, stability, and other properties critical to drug development.[20, 18, 19]

For polyprotic acids and bases, it is essential to consider the micro-pK_a values, which is the term of art for microscopic- or microstate- pK_a, of individual protonation and deprotonation sites. Micro-pK_a values refer to the pK_a values of specific sites on a molecule, rather than the overall pK_a value of the entire molecule. Knowing the micro-pK_a values of a polyprotic molecule can help us understand its behavior at different pH levels, and design drugs or other molecules with optimal properties.

The chemical space of small, "drug-like" molecules is vast, estimated to be approximately 10^{60}.[153] Experimental pK_a evaluation of all potential molecules is impractical, as only a modest number of reliable experimental pK_a values are available. As a result, researchers have developed various computational approaches to predict pK_a values, broadly classified into two categories: quantum mechanical (QM) and machine learning (ML) models.

QM models use different computational methods, such as density functional theory (DFT), semiempirical methods, or quantum mechanics/molecular mechanics (QM/MM) to compute the thermodynamics of protonation or deprotonation and thus the acid/base equilibrium. These methods are based on the principles of quantum mechanics and can provide accurate predictions of pK_a values, with root mean square errors (RMSEs) ranging from 0.6 to 1.6 pK_a units, particularly for related species, but they are computationally expensive.[45, 46, 47, 48, 49, 50]

ML models instead use machine learning algorithms, such as random forests (RF) and graph neural networks (GNNs), to predict pK_a values, based on training on either previously-computed or experimental data. ML models are less computationally expensive to evaluate than QM models, but they are generally not as accurate on novel compounds with RMSEs

ranging from 0.7 to 1.5 pK_a units. However, they are still useful for screening large numbers of molecules to identify potential drug candidates.[30, 31, 5, 6, 32, 33]

There are several pitfalls in designing new micro-pK_a models. The first one is to consider the correct tautomer, most likely to be the abundant form of the molecule in question at neutral pH in aqueous solution. This can lead to incorrect assignment of reaction sites in both the training and prediction steps of ML model development. Second, is the need to consider the molecule as a whole, including electronic and steric effects which may modulate the reactivity of functional groups and individual sites. The third is the lack of publicly available high-quality experimental pK_a datasets with enough data points to ensure generalization of the ML models and avoid overfitting. Although some models use private or commercial datasets,[5, 33, 31] a different approach, such as transfer learning,[154] is needed to overcome this challenge, until more experimental pK_a measurements become available.

In this work, we introduce a new method, **QupKake** (**Qu**antum **pKa** graph-neural-networ**K** **E**stimator), a model which combines graph neural networks with QM features from the GFN2-xTB semiempirical method,[54] for the prediction of small-molecule micro-pK_a. By combining both QM and ML we are able to achieve state-of-the-art micro-pK_a prediction accuracies, with RMSE between 0.5-0.8 pK_a units on experimental test sets, significantly lower than previous models. Additionally, QupKake is open-source with the intention of helping advance scientific research and discovery. The complete source code and all the datsets used in the training and testing of QupKake can be found on GitHub at: `https://github.com/hutchisonlab/QupKake`.

To achieve these results, we applied several techniques and methodologies. A three-step workflow includes a QM-based tautomer search step with an implicit water solvation model, enumeration of reaction sites using QM knowledge and graph neural networks, and a graph neural network based micro-pK_a prediction model that was trained using transfer learning. QupKake's unique design shows that combining the best of both worlds, QM and ML, can lead to better models and open up new possibilities for molecular design.

5.3 Methods

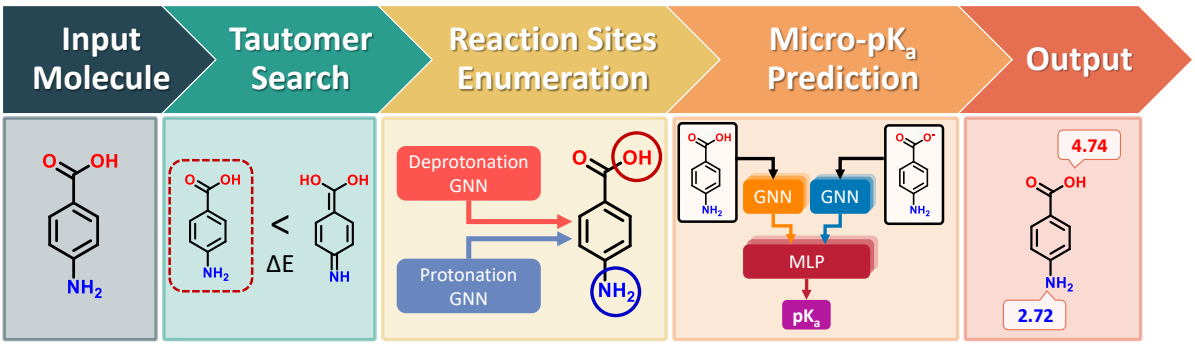


Figure 5.1: QupKake's workflow. The input molecule goes through three steps: tautomer search, reaction site enumeration, and micro-pK_a prediction. The output is the micro-pK_a value of each reaction site.

The workflow of QupKake has three main parts: tautomer search, reaction site enumeration, and micro-pK_a prediction. These steps are designed to ensure the accuracy of the predicted micro-pK_a values.

- **Tautomer search**: QupKake identifies the most stable tautomer of the input molecule. This is important because the micro-pK_a value of a site can vary depending on the tautomeric form.
- **Reaction site enumeration**: QupKake enumerates all of the possible protonation and deprotonation sites on the molecule. This takes into account the chemical structure of the molecule and the protonation states of its neighbors.
- **Micro-pK_a prediction**: QupKake predicts the micro-pK_a values of the enumerated reaction sites.

5.3.1 Tautomer Search

Certain molecules undergo tautomeric transitions, where they can exist in different forms due to the movement of protons within their structure. This proton shuffle is influenced by

factors like the solvent and typically leads to the dominance of one tautomer in a solution. These distinct tautomeric forms exhibit varying bonding arrangements, resulting in potentially different micro-pK_a values. Therefore, to understand the chemical and physical properties, it is essential to determine the most stable and prevalent tautomeric species within the solution.[155, 156, 157]

Although some machine learning models incorporate a tautomer search step, such as Schrödinger's Epik,[5] most do not.[32, 48, 33, 6, 31, 50] The identity of the most stable tautomer is a key factor for pK_a prediction, so incorporating a tautomer search step in an ML model should hypothetically yield more accurate results. Therefore, we employ the GFN2-xTB method to quickly identify the most stable tautomer.

The list of tautomers for each molecule was generated using the tautomer enumeration function of version 2022.03.3 of the RDKit software package.[69] For simplicity of the workflow, the tautomer search focused on neutral compounds. The total electronic energy of each tautomer was then calculated using version 4.6.1 of the GFN2-xTB method[54] with the analytical linearized Poisson-Boltzmann (ALPB) implicit solvation model in water.[158] The lowest energy tautomer, i.e., the most stable tautomer, was saved and used in the next steps, while the less stable tautomers were discarded. Consideration of tautomeric forms for acids, bases, and zwitterions is recognized as an area for future work, since the most stable tautomer of the neutral species may not be the most stable for other states.

5.3.2 Reaction Sites Enumeration

In the next stage of the QupKake workflow, we focus on enumerating potential reaction sites, specifically atoms with a higher probability of either gaining or losing a proton within the most stable tautomeric form identified in the previous step. Many of the currently available micro-pK_a models rely on predefined SMARTS patterns[159] to identify common acidic and basic groups. These patterns, however, can fail to consider other important molecular characteristics, such as the impact of neighboring groups or electrostatics on the reactivity of the reaction site. Furthermore, the use of different sets of SMARTS patterns by various models can lead to inconsistencies in the identification of reaction sites.

For this study, we used a tool called Conformer-Rotamer Ensemble Sampling Tool (CREST),[160] which employs GFN2-xTB calculations and has a protonation and deprotonation site screening function.[161] For protonation site screening, CREST identifies lone pairs and π orbitals as possible protonation sites, followed by geometry optimization and the ranking of protomers by their total GFN2-xTB energies. For the deprotonation sites screening, CREST iteratively removes protons from the input molecules, followed by geometry optimization and the ranking of protomers by their total GFN2-xTB energies.

We used the Protonation and Deprotonation Site Search option of CREST version 2.12[160] together with GFN2-xTB version 4.6.1. Because CREST tests all possible structures, it finds sites that are chemically unreasonable or rarely occurring, such as the protonation of aromatic carbons. This behavior has been reported previously and is attributed to a low activation barrier for proton transfer reactions predicted by GFN2-xTB, compared to other methods.[162, 54] Therefore, we constrained CREST to only output structures up to 10 eV higher in energy than the lowest energy structure. This reduces the number of possible protonation or deprotonation sites found by CREST to only the most stable reaction sites.

Our dataset consists of 1,475,879 molecules extracted from version 32 of the ChEMBL online database.[163, 164] The molecules were filtered to include only organic molecules (atoms H, C, N, O, S, P, F, Cl, Br, I) with ChemAxon[165] acidic or basic pK_a values between 0 and 14. In this work, acidic pK_a refers to the pK_a of a removal of an acidic proton, while basic pK_a refers to the protonation of a base. To further augment the dataset and teach the model about enantiomers, we converted each chiral molecule in the ChEMBL dataset to its enantiomer by inverting all chiral centers in the SMILES representation, recognizing that enantiomers intrinsically have the same pK_a values and features. This yielded an additional 376,202 molecules, for a total of 1,852,081 molecules. Molecular descriptor distributions of the dataset are shown in Figure D2 in Appendix D.

CREST protonation and deprotonation was then performed on each molecule in the dataset. CREST output structures were compared with the input molecule using openbabel version 3.1.1[166] in order to identify reaction sites. This was done separately for the protonation and deprotonation processes to generate a separate dataset for each reaction. Since CREST performs reactive molecular dynamics, molecular rearrangement or degradation can

occur after protonation or deprotonation, such as ring closure or splitting. Therefore, any molecule with a structural change other than an addition or removal of a proton was removed from the dataset. In total, there are 1,331,870 protonated molecules with a total of 2,265,676 protonation sites, and 1,214,117 deprotonated molecules with a total of 1,799,457 deprotonation sites.

Comparison of the protonation and deprotonation sites found by CREST versus those identified using SMARTS patterns is intended to provide insight into their agreement and to explore whether CREST identifies reaction sites beyond those captured by SMARTS patterns, and vice versa. For this purpose, we compared the reaction sites in our dataset with the SMARTS patterns compiled by *Pan et. al.*,[33] encompass 54 acidic group patterns and 89 basic group patterns. This comparison with SMARTS patterns from previous literature is not to consider them as a "gold standard," but rather to evaluate how CREST's findings align with established patterns in the literature and to highlight any novel insights CREST might provide.

Overall, CREST and the SMARTS patterns agreed on 53.08% of the protonation sites and 72.32% of the deprotonation sites. This indicates that CREST results and the SMARTS patterns generally agree on the location of protonation and deprotonation sites. However, there are also some significant differences.

For protonation, 40.67% of the sites were found using the SMARTS patterns but not with CREST, while 6.24% of the sites were found with CREST but not with the SMARTS patterns. This discrepancy underlines that while CREST may identify some unique protonation sites not captured by the SMARTS patterns, due to its QM calculations that consider non-local and non-bonding interactions, the SMARTS patterns, on the other hand, tend to identify a broader array of protonation sites without regard to the entire molecule or relative energies. This results in a significant number of potential sites, including some higher in energy and thus a decreased agreement between the two methods. For deprotonation, the percentages are slightly closer, with 24.67% of the sites found with the SMARTS patterns but not with CREST, and 3.31% found with CREST but not with the SMARTS patterns. This suggests that CREST and the SMARTS patterns are more similar in their identification of deprotonation sites, but the SMARTS patterns tend to find more possible deprotonation

sites. This comparison does not expect a complete concurrence between SMARTS patterns and CREST, particularly given the limitations of SMARTS in capturing complex QM interactions. While deriving tautomerization rules from quantum chemical calculations has been performed successfully,[156, 157] as of yet, we have not observed the same for protonation or deprotonation rules. We hope that this work and data set can help to provide a basis for similar efforts.

Overall, a comparison of the CREST results and the SMARTS patterns reveals that they generally agree on the location of protonation and deprotonation sites, albeit with differences. Several examples of the different predictions of protonation and deprotonation sites are shown in Figure D8 and Figure D9 in Appendix D, respectively.

Again, while the SMARTS patterns may be useful, we focus on CREST reaction sites because it uses GFN2-xTB calculations to rank the relative thermodynamics of protonated and deprotonated molecules by total electronic energy, which will be useful later. However, its exhaustive approach involving meta-dynamics and molecular dynamics (MD) simulations contributes to its relatively slow performance, taking up to several hours for each molecule. Therefore, an ML model trained on the CREST dataset can be used as a surrogate for CREST reaction site determination to provide fast, high-accuracy predictions.

5.3.2.1 Graph Neural Networks Models

Machine learning methods have been successfully applied to predict molecular properties in recent years,[167, 168, 169, 170, 171, 172, 173, 174, 175] including macro- and micro-pK_a.[32, 33, 5, 50, 31] Graph neural networks (GNNs) are a type of ML model that can be used to learn representations of graphs, where nodes represent atoms and edges represent bonds. GNNs have been shown to be effective for molecular property prediction tasks.[33, 30, 5, 176, 177, 178, 179, 180, 181] In this study, we constructed and trained two GNNs on the CREST dataset to predict protonation and deprotonation sites. We found that our GNN models outperformed CREST in terms of prediction accuracy and speed. Our results suggest that GNNs can be used to accelerate QupKake's workflow by identifying potential reaction sites more efficiently.

We generate a molecular graph representation of each molecule using version 2.3.0 of the *PyTorch Geometric*[182] package. For each atom (node), bond (edge), and molecule (graph), we generated a set of features using RDKit and GFN2-xTB. For a complete list of features, see Table D1 in Appendix D. Our protonation and deprotonation site prediction models are node-level prediction models. The input to each model is a graph representation of a molecule in the shape of $(61, N)$, where N is the number of atoms. The output of each model is a one-dimensional binary vector in the shape of $(1, N)$, where each element of the vector indicates whether the corresponding atom is a predicted protonation or deprotonation site, respectively.

We split the CREST dataset into train, validation, and test sets in an 80:10:10 ratio, respectively. We trained the models using using *Python* version 3.9, *PyTorch*[183] version 2.0.0 and *PyTorch Lightning*[184] version 2.0.2. We tested three different GNN architectures: Graph Convolution Network,[185] Graph Attention Network,[186] and Graph Transformer Network.[187] We used *Optuna*[188] version 3.2.0 to find the best architecture and hyperparameters for the models, with the goal of maximizing accuracy. Early stopping of the model training was used to prevent overfitting.

To confirm that the GFN2-xTB features contribute significantly to the models' performance, we used the *Integrated Gradients* algorithm,[189] as implemented in version 0.6.0 of the *Captum* Python package,[190] to quantify the importance of each feature in predicting the reaction site. The Integrated Gradients algorithm is a model-agnostic attribution method that can be used to explain the predictions of any machine learning model. It works by computing the gradient of the model's output with respect to its input features, and then integrating the gradients over a path from a baseline input to the actual input. The resulting values represent the importance of each feature in contributing to the prediction of the model.

5.3.3 Micro-pK_a Prediction Model

The final step in QupKake's workflow is the prediction of the micro-pK_a values for the reaction sites previously predicted. The model architecture follows a similar schematic to

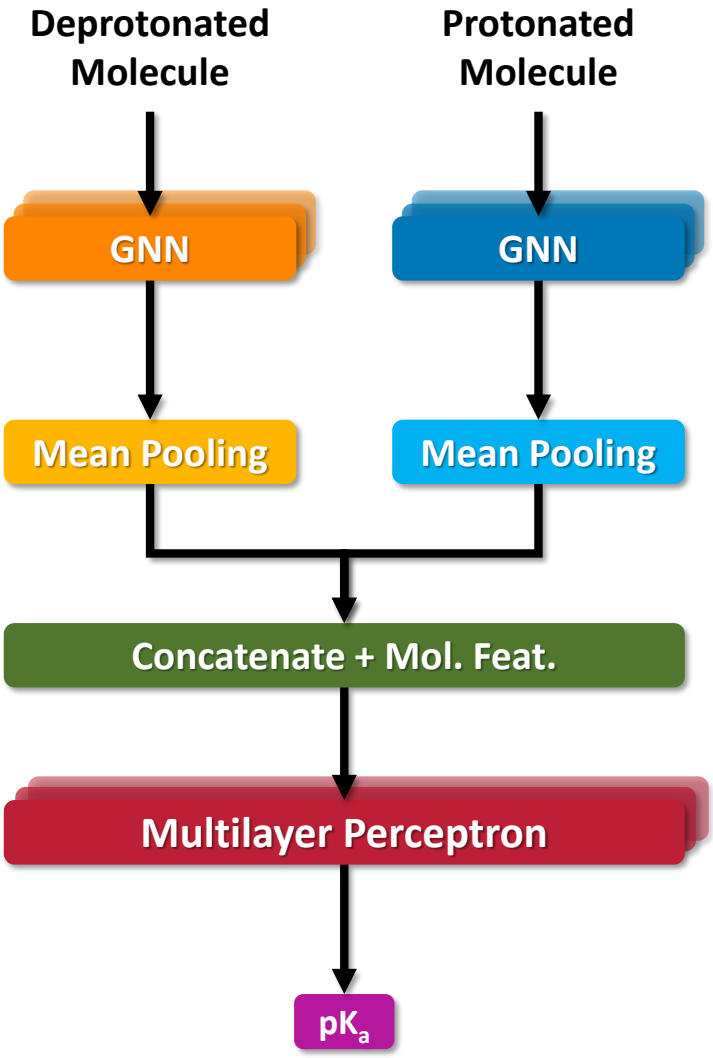


Figure 5.2: The simplified micro-pK_a model architecture. The model takes in two input molecules, where one molecules is the protonated version of the other. The model's output is the pK_a value of the protonation\deprotonation reaction between the two species. See Figure D13 in Appendix D for the full model architecture.

the one proposed by *Mayr et. al.*,[6] in which the input consists of a graph representation of both the original input molecule and its protonated or deprotonated version, depending on the reaction site. Each input molecule undergoes several GNN layers, followed by a global mean graph pooling layer. The outputs are then concatenated, together with the molecular features vector of each input molecule, into a one dimensional vector. This vector is passed to several linear layers that outputs the pK_a of that site (Figure 5.2). The implementation of this model used the same tools and packages as described above.

A comprehensive and high-quality public micro-pK_a datasets, that would provide enough diverse experimental values to train a model are hard to come by. While the iBond database[191] contains over 30,000 equilibrium pK_a values in aqueous and non-aqueous solvents, it is not readily accessible and encompasses pK_a values across 46 different solvents, complicating its direct application for model training due to solvent diversity. Some models, such as Schrödinger's Epik,[5] use proprietary data that are not accessible to others, inhibiting the progress of scientific discovery. The Czodrowski group has curated a dataset of $\sim 6,000$ organic compounds with experimental pK_a values[32] and used ChemAxon's Marvin[165] to find the reaction site. Molecular descriptors and the distribution of pK_a valuesfor the experimental dataset can be found in Figures D3 and D4 in Appendix D. However, 6,000 experimental data points are not enough to adequately train a GNN model.[192] Furthermore, during model development, we have found that the experimental dataset requires augmentation and cleaning as some molecules had to be neutralized, had a chemically improbable assignment of the reaction site, or had calculation errors with GFN2-xTB. This narrowed the experimental dataset to 5,637 compounds.

To mitigate this problem, the model was trained on the previously mentioned ChEMBL dataset, consisting of ~ 2.5 million predicted acidic and basic pK_a values over ~ 1.5 million molecules. As the molecules in this dataset have been previously processed by CREST, the lowest energy protomer or deprotomer was assigned to the predicted pK_a value. That is, Marvin's most basic pK_a prediction was assumed to describe the protonation reaction of the molecule and the most stable protomer found by CREST. The same applies to the most acidic pK_a and the molecule's most stable deprotomer. The assignment of predicted macro-pK_a value to a reaction center and treating it as a micro-pK_a value is because in most cases

CREST found only one acidic or basic reaction center that was energetically accessible per molecule, which is in agreement with our assumption. In cases where there were multiple reaction centers, this assumption can be flawed, but it is how we were able to achieve a diverse training set. As before, the model was divided into training, validation, and testing sets in an 80:10:10 ratio, respectively.

Transfer learning was then used to fine-tune the model with the 5,637 experimental micro-pK_a values, which was randomly split to 80:20 ratio of training and validation sets. Transfer learning has been a widely used technique in the ML world for fine-tuning a pre-trained model on new information with the purpose of increasing accuracy or improving performance on another related task.[193, 194, 195]. In this case, the pre-trained model was able to take advantage of its existing knowledge of molecular structure and properties to learn to predict micro-pK_a values more accurately.

To test the performance of the model and compare it to other available models, we used two public datasets also curated by the Czodrowski group.[32] Those datasets consist of 279 molecules from the Novartis drug company, and the other dataset consist of 123 molecules with experimental pK_a values from different literature sources. Molecular descriptors and the distribution of pK_a valuesfor the experimental dataset can be found in Figures D5 and D6 in Appendix D. The model's performance was also compared with the datasets of Statistical Assessment of Protein and Ligand Modeling (SAMPL) in public challenges of prediction SAMPL6,[2] SAMPL7,[3] and SAMPL8[4] micro-pK_a prediction public challenges.

5.4 Results and discussion

5.4.1 Reaction Sites Enumeration

5.4.1.1 Model Performance

The tuned and trained micro-pK_a prediction models were evaluated on an out-of-sample set of 133,188 molecules with protonation sites and 121,413 molecules with deprotonation sites, randomly selected from the CREST datasets. The models achieved very high accu-

racy, with 98.2% and 98.8% accuracy for protonation and deprotonation site enumeration, respectively. These results are shown in Figure D10. The optimized hyperparameters are listed in Table D2 in Appendix D.

5.4.1.2 Feature Importance

To better understand which features are the most important for the protonation sites enumeration model, we performed atom and bond feature importance analysis as described in section 5.3.2.1 (Figure D11). The most important atom feature is "Atom Type: N", which indicates whether the atom is a nitrogen atom. This makes chemical sense, as nitrogens are more likely to be protonated than other types of atoms in neutral molecules. Furthermore, 96% of the protonated atoms in the CREST dataset are nitrogens (Figure D7a), which supports the importance of this feature.

The next two most important atom features are the covalent coordination number ("xTB Coord Number") and the susceptibility to radical attack Fukui(0) index ("xTB Fukui(0)"), both of which are calculated using GFN2-xTB. This shows that QM features contribute significantly to the model's performance.

The most important bond feature is "Bond Type: SINGLE", followed by the "Wiberg Bond Order", which is calculated from the GFN2-xTB results. The computed bond order is a continuous real value based on electron density, and can thus indicate variations in bond strength.[196] This corroborates the importance of QM features, as well as the importance of single bonds in protonation reactions.

Overall, the feature importance analysis shows that the model is able to leverage both topological and quantum mechanical features to accurately predict protonation sites. This is an important finding as it demonstrates that the model can be used to predict protonation sites for a wide range of molecules, even those for which experimental data is not available.

Similarly, to better understand which features are the most important for the deprotonation sites enumeration model, we performed atom and bond feature importance analysis (Figures D12).

The most important atom feature is "Is HBD", which indicates whether the atom is

a hydrogen-bond donor. This makes chemical sense, as hydrogen bond donors are more likely to be deprotonated than other types of atoms. The highest GFN2-xTB feature, the atom's partial charge ("xTB Partial Charge"), is only in seventh place, indicating that quantum mechanical (QM) features are less important for deprotonation site prediction than for protonation site prediction.

Similarly to the protonation site enumeration model, the most important bond feature is "Bond Type: SINGLE". However, the GFN2-xTB calculation "Wiberg Bond Order" feature is only the fifth most important feature, again indicating that QM features are not as important for deprotonation site prediction as for protonation site prediction.

Overall, the feature importance analysis shows that the deprotonation sites enumeration model is less reliant on QM features than the protonation sites enumeration model.

Despite the lower importance of QM features for deprotonation site prediction, the model is still able to achieve high accuracy. This is likely because the model is able to learn complex relationships between the topological and chemical properties of the molecule.

5.4.2 Micro-pK_a Prediction Model

5.4.2.1 Model Performance

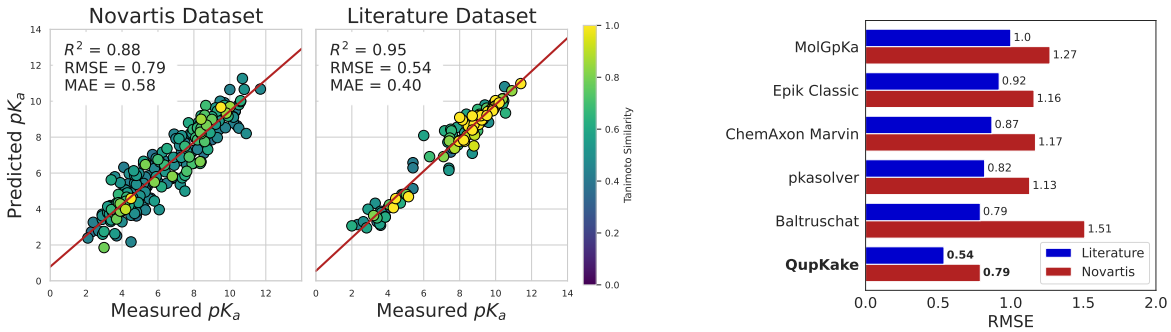
The tuned and trained micro-pK_a prediction model was validated on five external test sets: the Novartis dataset (containing 280 molecules), the Literature dataset (containing 122 molecules), and the SAMPL6, SAMPL7, and SAMPL8 datasets (containing 24, 20, and 21 molecules, respectively). The results of this evaluation are shown in Figures 5.3a and 5.3c. The optimized hyperparameters are listed in Table D3 in Appendix D.

On the Novartis and Literature test sets, the model achieved low prediction errors with root mean square errors (RMSEs) of 0.79 and 0.54 pK_a units, respectively, and mean absolute errors (MAEs) of 0.55 and 0.39 pK_a units, respectively. The model also achieved high coefficients of determination (R^2) for both test sets, with values of 0.88 and 0.95, respectively. These results demonstrate that the model is able to accurately predict micro-pK_a values for a wide range of organic molecules.

To put QupKake's high performance into context, Figure 5.3b shows a comparison of the

(a)

(b)



(c)

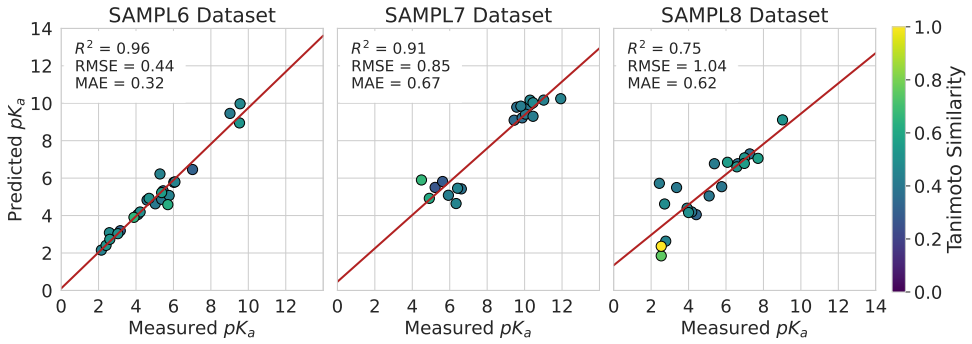


Figure 5.3: Micro-pK_a predictions versus the measured micro-pK_a values of the a Novartis dataset and Literature dataset, as well as the c SAMPL6, SAMPL7 and SAMPL8 datasets. Data points are colored according to the highest Tanimoto similarity score of the molecule in the test set versus the molecules in the experimental training set. The best-fit linear regression line is shown in red. b RMSE comparison of the Novartis and Litareture datasets between QupKake and five other models. The RMSE values for the five other models were obtained from *Mayr et al.*.[6]

RMSE of QupKake versus five other models on the Novartis and Literature datasets. Those models include MolGpKa,[33] Schrödinger's Epik Classic,[197] ChemAxon's Marvin,[165] pkasolver, [6] and the Baltruschat model from the Czodrowski group.[32] The RMSE values were obtained from *Mayr et al.*.[6] It is clear to see that QupKake significantly outperforms all other models, with RMSE differences of 0.34 and 0.25 pK_a units on the Novartis and Literature test sets between QupKake and the next-best models.

The Novartis and Literature test sets were both obtained from the Czodrowski group,[32] which stated that the test sets and the experimental training set do not have the same molecules, which could lead to the model "memorizing" values instead of learning. However, we have found that this is not the case, since several molecules appear in both these test sets and the experimental training set. To see if the model actually memorized the molecules and their pK_a values, or whether QupKake was able to generalize, we removed any molecules from the test sets with high Tanimoto similarity scores (< 0.8) and compared the performance of the model with the rest of the molecules (Figure D15). Tanimoto similarity scores are defined as the ratio of the intersection of the two sets of fingerprints over the union of the two sets and have been widely used to calculate molecular similarities in various applications.[198] The R^2, RMSE, and MAE of the Novartis dataset remained the same, while the RMSE and MAE of the Literature dataset slightly increased to 0.59 and 0.43, respectively, while R^2 was unchanged. These results show that the model did not simply memorize values and was able to learn and generalize on a range of different species and pK_a values. Furthermore, we have not found any correlation between the Tanimoto similarity score and the pK_a error (Figure D14), again suggesting the QupKake model has strong generalization.

To illustrate the benefits of transfer learning in refining the model's accuracy, we conducted experiments with two distinct models without transfer learning. The first is the initial model, trained solely on the ChEMBL dataset without fine-tuning using the experimental data, exhibited an increase in the RMSE for the Novartis and Literature test sets to 1.09 and 0.86, respectively, compared to the fine-tuned model (Figure D17). The second model, trained exclusively on the experimental data with identical hyperparameters to those of the fine-tuned model, demonstrated a more pronounced increase in RMSE for the Novartis and Literature test sets, reaching 1.79 and 1.31, respectively (Figure D18). These results affirm

our hypothesis that transfer learning markedly improves the model's performance.

Table 5.1: Comparison of QupKake's accuracy versus the top ranked submissions[1] in the SAMPL6,[1, 2] SAMPL7[3] and SAMPL8[4] pK_a prediction challange. The table is sorted from lowest to highest RMSE of the models in each SAMPL challenge. While the Epik 7 Ensemble model[5] was not submitted to the SAMPL6 challenge, we included it here as it is the most recently published micro-pK_a prediction model, as well as for providing its performance on the SAMPL6 dataset.

SAMPL6				**SAMPL7**				**SAMPL8**			
Model	**RMSE**	**MAE**	**R^2**	**Model**	**RMSE**	**MAE**	**R^2**	**Model**	**RMSE**	**MAE**	**R^2**
QupKake	**0.44**	**0.32**	**0.96**	EC_RISM	0.72	0.53	0.93	**QupKake**	**1.04**	**0.62**	**0.75**
Epik 7 Ensemble	0.61	0.48	0.95	**QupKake**	**0.85**	**0.67**	**0.91**	DeeepGP	3.17	2.62	0.15
Grimme	0.68	0.58	0.94	IEFPCM/MST	1.82	1.30	0.56	3DS	3.44	2.49	0.27
S+pK_a	0.73	0.59	0.93	DFT_M05-2X_SMD	2.90	2.28	0.03	ChemAxon	4.18	2.82	0.09
ACD/pK_a Classic	0.79	0.56	0.92	TZVP-QM	2.90	2.75	0.23	ECRISM	4.56	3.05	0.18
COSMOtherm pK_a	0.90	0.71	0.90	Gaussian Process	3.49	2.91	0.30	ZhiyiWu	4.73	3.37	0.05
MoKa	0.94	0.77	0.88	DFT_M06-2X_SMD	5.12	2.56	0.20	uESE_extra	6.80	5.33	0.09
Epik Classic	0.95	0.78	0.91	Gaussian corrected	5.36	5.12	0.76	—	—	—	—

Despite being trained on the ChemAxon dataset, QupKake outperforms the Marvin program in identifying reaction sites that correlate more closely with experimental pK_a values in the Novartis and Literature datasets. This indicates that QupKake's ability to identify accurate reaction sites generalizes well beyond the training data. Even when evaluated using the reaction sites identified by Marvin, QupKake still surpasses the performance of other micro-pK_a prediction models. On the Novartis and Literature test sets, QupKake achieves RMSEs of 1.00 and 0.59 pK_a units, respectively (Figure D16). These RMSEs are higher than those obtained using QupKake's own reaction sites, but they remain lower than those of the other five models compared in Figure 5.3b. This demonstrates QupKake's robustness and ability to provide accurate micro-pK_a predictions even when using reaction sites identified by external tools.

The model was also tested on the SAMPL6, SAMPL7, and SAMPL8 pK_a prediction challenges (Figure 5.3c). QupKake outperformed all of the submitted models for the SAMPL6[1, 2] and SAMPL8 models,[4] and would have been ranked first if the challenges were still open

for submission. Table 5.1 shows the superior performance of QupKake on the SAMPL6 and SAMPL8 datasets compared to the best performing submissions, as well as the latest Epik 7 Ensemble model from Schrödinger.[5] QupKake performs slightly worse on the SAMPL7 dataset and would be ranked second by RMSE compared to the other submissions.[3]

Beyond ranking the performance on test sets, the model should be evaluated for trends in the most accurate and least accurate micro-pK_a predictions to consider potential chemical motifs and model bias on certain acidic or basic groups. Figures D19 and D20 in Appendix D show the 20 most accurate and least accurate micro-pK_a predictions on the Novartis dataset. No clear pattern is observed, suggesting that there is no noticable bias of the model. Although there are several examples of poor predictions on the acidic pK_a of amides in D20, it is not a significant trend across the entire testing set. The RMSE of only acidic amides in the Novartis dataset is 0.94 pK_a units, which, while slightly higher than the RMSE of the entire set (0.79), does not indicate clear bias against these groups.

Additionally, Thapa & Raghavachari compiled a set of organic molecules categorized by functional group. They calculated pKa values using high-level QM methods, employing both implicit and explicit water solvation models. Their results were tabulated for each group [199]. By applying the QupKake model to each functional group list, we can assess whether QupKake exhibits differential pKa prediction accuracy across groups (Tables D4 – D15). As anticipated, QupKake demonstrates higher performance on functional groups well-represented within the training set. These include nitrogen-containing aromatic heterocycles, primary and secondary amines, carboxylic acids, anilines, and benzoic acids. On the contrary, groups less prevalent in the training data, such as aliphatic alcohols & thiols, phenols, and thiophenols, yielded lower precision. Notably, "carbon acids" (i.e., deprotonation of aliphatic carbon atoms) were absent from the training set, making the current model unable to predict the reaction site for this group.

As with any ML model, the accuracy and precision of the model output is directly related to the accuracy and precision of the training data. As mentioned before, high-quality micro-pK_a data is hard to obtain. Due that, QupKake is trained only on one pK_a value per compound of a relatively small section of the vast chemical space. Therefore, we acknowledge that we can only have high confidence in the most acidic or basic micro-pK_a values that

QupKake predicts, while we have less confidence in the micro-pK_a predictions of additional reaction sites in compounds with multiple sites (e.g., polyprotic acids). Further experimental pK_a data, for example from automated characterization would greatly improve the accuracy of future models.

Overall, the results of the external evaluation demonstrate that the tuned and trained QupKake prediction model is a highly accurate and reliable tool for predicting micro-pK_a values for organic molecules. The model outperforms all other state-of-the-art models on a variety of benchmark datasets and is able to generalize to new data, including examples which are more challenging than the training data. This suggests that the model could be used to predict micro-pK_a values for a wide range of organic molecules, including drug candidates and other molecules of interest.

5.4.2.2 Feature Importance

As with the reaction site enumeration models, feature importance analysis can give useful insights into which features contribute to the the micro-pK_a prediction model, such as whether including QM features improves the model performance. We performed a feature importance analysis on the atomic, bond, and molecular features using the *Integrated Gradients* algorithm as described earlier. However, as the micro-pK_a prediction model's architecture (Figure D13) is more complex than the reaction site enumeration models, as well as being a graph regression model, compared to a node classification model in the case of the site enumeration models, feature importance analysis is less straightforward.

In the micro-pK_a model architecture, the atomic and bond feature vectors of each atom and bond in both the protonated and deprotonated molecules are first passed through several GNN layers, which are then pooled into a one-dimensional vector. It is then concatenated with the molecular feature vectors and passes through several linear layers, which output the predicted micro-pK_a. As the feature vectors, especially the atomic and bond features, have gone through several transformations, it can be difficult to deduce how and why each feature affected the model's performance.

Figure D21a shows the importance of the atomic features, with "Atom Type: N" having

the highest score. We hypothesized that this is due to the abundance of nitrogen reaction sites found by CREST, and therefore by the reaction site GNNs. The next important atomic feature is the "xTB Alpha", which is the atomic polarizability calculated by GFN2-xTB. This indicates that QM features provide useful information to the model and improve its performance. The next features by importance score include "Is HBA", "Formal Charge: 0", and "Atom Type: O", which, as before, might indicate their abundance in the dataset.

Similarly, the most important bond features shown in Figure D21b are "Is In Ring" and "Is Conjugated", indicating the high prevalence of aromatic rings in the dataset. The "Wiberg Bond Order", calculated by GFN2-xTB, is the third most important feature, again proving that QM features contribute to the model performance.

The molecular features did not pass through the GNN layers, and thus should have a more direct and interpretable impact on the model. Figure D21c shows that the five RDKit features, "RadiusOfGyration", "Eccentricity", "Spherocity", "FractionCSP3", and "Asphericity", have almost identical importance scores, while the protonation energy ("xTB-Energy", the energy difference between the protonated and deprotonated molecules) have a very low score. This could indicate that there is a negligible correlation between GFN2-xTB energies and pK_a s, while the RDKit features contribute similar information to the model. Improved semiempirical methods with better treatment of solvation effects,[200] or ML-based models[201] may improve the influence of these molecular features.

5.4.3 Model Speed

Although GFN2-xTB is a relatively fast semiempirical method, especially compared to higher-level methods,[53] it is significantly slower than using RDKit alone to calculate graph-level features. As with many things in life, however, there is an inverse relationship between speed and accuracy and the QM features prove to be important in both tautomer selection and the ML models.

To evaluate how fast it takes for a molecule to pass through QupKake's workflow, we performed a benchmark test using a server running a 3.85GHz AMD EPYC 9374F CPU with 32 cores with a shared memory framework. Although ML tasks generally run faster

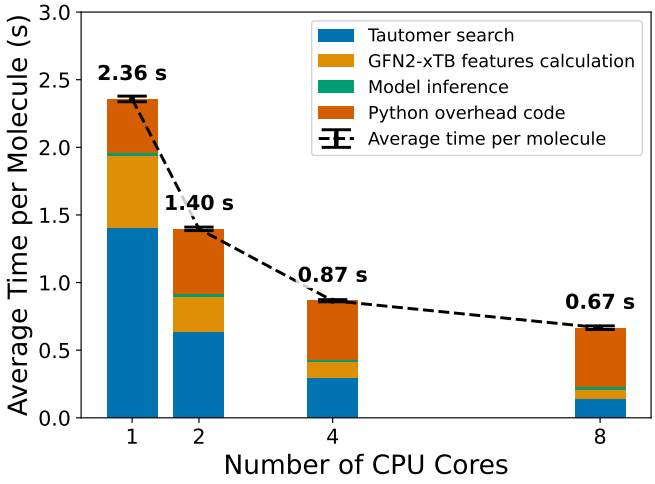


Figure 5.4: Average compute time per molecule across the 280 molecules in the Novartis test set as a function of the number of CPU cores, indicating time spent in the tautomer search calculations, GFN2-xTB feature calculations, ML model inference, and Python overhead from the QupKake model code, including RDKit descriptors. The error bars show the standard deviation of the average compute time per molecule over 10 trials.

on GPU, the rate-limiting steps in QupKake are the GFN2-xTB calculations, which do not take advantage of GPU acceleration. However, GFN2-xTB can utilize parallel processing on multiple cores to increase its calculation speed. Additionally, to achieve even better better performance, we utilized Python's multiprocessing module to parallelize the dataset's preprocessing.

Using the 280 molecules in the Novartis dataset as our benchmark, we measured QupKake's average compute time per molecule from start to finish, as well as how long each step took, as a function of the number of CPU cores (Figure 5.4). We repeated this 10 times to minimize random events that can skew the benchmark timings. The very small standard deviations, around 0.02 seconds per molecule in the single CPU core case, show that QupKake's execution time is consistent. Of course, some variance is expected when different molecular sets are used.

It is clear that the tautomer search step, which uses GFN2-xTB to find the most stable tautomer, takes a significant time to compute, followed by the calculations of the GFN2-xTB features for the reaction site search and the micro-pK_a prediction steps. As mentioned before, the actual model inference compute time is negligible, even when using a CPU.

As more CPU cores were used in parallel, the compute time for the tautomer search and the GFN2-xTB calculations decreased inversely at an approximately linear rate. That is, the average compute time for those steps using four CPU cores is approximately half the compute time using two cores and about a quarter compared to using a single core (Figure D22). These steps run well in parallel because multiple tautomers can be calculated at once, and many components of a quantum calculation such as GFN-xTB also run well in parallel.

In contrast, the compute time for the overhead Python code that could not be parallelized remained approximately the same regardless of how many CPU cores were used. Therefore, the overall speedup achievable by using more CPU cores is limited as it does not scale linearly with the number of cores (Figure D23). Thus, using more than two to four cores per molecule provides only a minor improvement in speedup.

In general, the use of multiple CPU cores can be a valuable tool to improve QupKake's performance. However, the benefits of using multiple cores are not linear, and the overhead of

using multiple cores can also be a factor. Therefore, in common cases such as the evaluation of multiple compounds in a set, it is more effective to run separate QupKake calculations in parallel rather than dedicating many cores to each.

5.5 Future Directions

We have shown that combining semiempirical QM features with ML improves the micro-pK_a predictions for small "drug-like" molecules. However, while GFN2-xTB is a relatively fast method, especially compared to higher-level methods,[53] it is still slower than "pure" ML methods that use only RDKit features.[5, 33, 6] As described above, the greatest bottlenecks in QupKake's workflow are the GFN2-xTB calculations, and finding a faster replacement that provides similar, or better accuracy can improve future versions of QupKake.

For example, the recently published MolTaut model[202] which uses a GNN to rank tautomer stability in aqueous solutions, could replace the current tautomer search step, which uses GFN2-xTB calculations to do so. Other models, such as Auto3D or AIMNet2,[203, 201] which uses a message-passing approach, can also be used to calculate the relative energies of the tautomers.

ML model can also be used to predict certain atomic and bond features, which can make the use of GFN2-xTB calculations obsolete. Features such as atomic partial charges,[204, 205] Fukui indices,[206] and bond orders[207] already exist and can be integrated with future iterations of QupKake. Other GFN2-xTB features that prove important for micro-pK_a predictions, such as atomic polarizabilities and coordination numbers, currently do not have published models. However, it is possible to build surrogate models to predict these values, in a manner similar to QupKake's reaction site models, training on the calculated values from GFN2-xTB or a higher-level method.

Caldeweyher et al. recently introduced an alternative method for calculating various QM features, including coordination numbers, atomic polarizabilities, and partial charges [208]. Their work presents kallisto, a program that employs equations parameterized to GFN2-xTB data to compute these features. While corrections to GFN2-xTB atomic polarizabilities

have been noted [209], this approach offers the potential to significantly accelerate feature calculation to the level of GFN2-xTB computational efficiency, without directly utilizing the GFN2-xTB method.

5.6 Conclusions

In this work, we have presented QupKake, a novel and effective workflow for predicting micro-pK_a values of small organic molecules. QupKakeleverages the power of graph neural networks (GNNs) and semiempirical quantum mechanical (QM) features, namely the GFN2-xTB method, to achieve exceptional accuracy and generalization. Our comprehensive evaluation demonstrates that QupKake outperforms all other state-of-the-art models, yielding low prediction errors on five external test sets, with RMSEs between 0.5-0.8 pK_a units.

Further analysis of QupKake's feature importance revealed the crucial role of QM features, such as the coordination number and Wiberg bond order, in both reaction site enumeration and micro-pK_a prediction models. Additionally, topological features, including atom and bond types, were also found to be essential for the model's performance.

While QupKake exhibits remarkable accuracy and generalization, we also investigated its speed and identified the tautomer search and GFN2-xTB calculations as the most time-consuming steps in the workflow. To address this challenge, we have outlined several promising research directions, including developing a faster replacement for GFN2-xTB calculations and utilizing ML models to predict certain atomic and bond features.

We believe that QupKake represents a significant contribution to the field of computational chemistry, offering a powerful tool for predicting micro-pK_a values of organic molecules. Its potential applications span a wide range of fields, including drug discovery and materials science. Moreover, the use of transfer learning, using abundant computed predictions to train an initial model, followed by experimental refinement, offers a clear mechanism to improve model accuracy in chemistry, when accurate experimental data may be scarce.

6.0 Conclusions and Future Directions

This work has explored the combination of quantum mechanical calculations together with machine learning and genetic algorithms to design and discover new materials with desirable electronic and molecular properties. Using these advanced computational techniques, the research has shown significant potential for accelerating material and drug discovery. One focus was on π-conjugated polymers, which have applications in organic electronics, such as solar cells, transistors, and light-emitting devices. Furthermore, this research included micro-pK_a predictions for drug-like molecules, which are crucial to understanding the behavior of pharmaceutical compounds in biological systems.

QM methods, such as DFT, were used to provide information on the electronic structure of materials. Despite their accuracy, these methods are computationally intensive. To mitigate this, semi-empirical methods, mainly GFN2-xTB, were introduced as a viable alternative, offering faster calculations with some trade-offs in accuracy. GFN2-xTB, for example, has been shown to provide reliable geometries and approximate electronic properties, making it suitable for high-throughput screening of large molecular datasets in the selection step of a Genetic Algorithm (GA), as well as providing important features for ML applications. This balance between speed and accuracy is essential for practical applications where large-scale screening is necessary.

One of the key advantages of ML in molecular design is its ability to quickly predict properties of new unseen compounds. This capability is expecially important for discovering materials with novel properties that are not present in the training datasets. The combination of QM and ML allows for the accurate prediction of material properties with significantly reduced computational cost compared to QM calculations alone. The ability to integrate different types of data, such as structural features and quantum mechanical properties, into predictive models further enhances the resilience and applicability of these techniques.

The integration of QM methods with ML further enhanced the efficiency of property predictions. In Chapter 2, a random forest ML model was trained on a dataset of tetramers and hexamers with calculated reorganization energies, using geometrical features calculated

using GFN2-xTB. This method has been shown to accurately identify candidates with low reorganization energies, achieving a $\sim$13$\times$ speedup. In Chapter 5 of this work, we introduced QupKake, a graph neural network (GNN) based model that uses GFN2-xTB in multiple steps of its workflow, including finding the lowest energy tautomer and calculating the atomic, bond and molecular features for the micro-pK_a prediction model.

In Chapter 3, we identified that a low HOMO-LOMO gap of an oligomer is a predictor for a stable triplet ground state. This gave us the basis in which we built upon in Chapter 4, where a GA was used to optimize the exploration of a large chemical space. The GA efficiently searched for molecules with a low HOMO-LUMO gap, using GFN2-xTB calculations in the selection step. It successfully found 1,400 possible candidates out of 1.5 million potential monomer combinations. The GA also helped identify certain design rules that can guide future searches, such as specific monomers and the use of a vinyl bridge.

The integration of GAs with QM and ML methods provided a robust approach to material discovery, enabling the identification of new materials with exceptional properties. This combination allows for an efficient exploration of the chemical space, where QM methods ensure the accuracy of property evaluations or provide important features, ML models provide rapid predictions, and GAs optimize the search process. This integrated approach was demonstrated through several case studies on the reorganization energies and the stability of a triplet ground state of π-conjugated polymers, as well as the micro-pK_a values of drug-like molecules, which showcase the practical implementation and effectiveness of the proposed methodology. The successful application of these methods highlights the potential for computational techniques to transform material discovery.

6.1 Future Directions

The integration of quantum mechanical methods, machine learning, and genetic algorithms presents numerous opportunities for future research. One promising direction is the expansion of the chemical space explored in this thesis. While the current work focused on π-conjugated polymers and drug-like molecules, the methodologies developed can be ex-

tended to other classes of materials, including monomers that were not used in this work and molecules and polymers with pK_a values outside of the biological range. By broadening the scope of materials studied, researchers can uncover new applications and functionalities that were previously unexplored, such as understanding their degradation in water as a result of protonation or deprotonation.

Another important future direction is the improvement of machine learning models to handle more complex and diverse datasets. One hurdle encountered in Chapter 5 was the lack of publicly available large enough datasets of experimental pK_a values. As the availability of experimental and computational data continues to grow, there is a need for more sophisticated ML algorithms that can learn from these vast datasets and make accurate predictions. Techniques such as transfer learning, as it has been utilized in this work, and active learning, where models iteratively query new data points to improve performance, hold great promise in this regard. Automated data collection, such as the use of robotic labs to perform measurements, can be an efficient way to greatly increase the number of data points for ML applications.

Furthermore, while semi-empirical methods such as GFN2-xTB have shown their successful implementation in ML and GA workflows, using ML models as surrogates for semi-empirical features can accelerate the search for better materials and make it more efficient. For example, a ML model can be trained to predict the HOMO-LOMO gap of oligomers such as those used in Chapters 3 and 4. This model can then be used in a GA, instead of the current implementation with GFN2-xTB, to potentially accelerate the search speed and broaden the search space of new polymers with a stable triplet ground state. ML models can also be trained on specific atomic or bond characteristics, such as partial charges and bond orders, for integration into GNN models such as QupKake.

Finally, the development of more efficient and accurate semi-empirical methods remains a critical area of research. While methods like GFN2-xTB provide a good balance between speed and accuracy, there is still room for improvement. Advancements in this field could lead to even faster and more reliable calculations, enabling the exploration of larger chemical spaces and more complex systems. Combining these improved semi-empirical methods with ML and GAs could further enhance the efficiency of material discovery and optimization,

paving the way for the development of next-generation materials with tailored properties for a wide range of applications.

Appendix A Machine Learning to Accelerate Screening for Marcus Reorganization Energies

A.1 Code and Data Availability

Full code, data files, and analysis notebooks are available at `https://github.com/Shualdon/ReorganizationEnergy`

A.2 Supplementary Information

(a) (b)

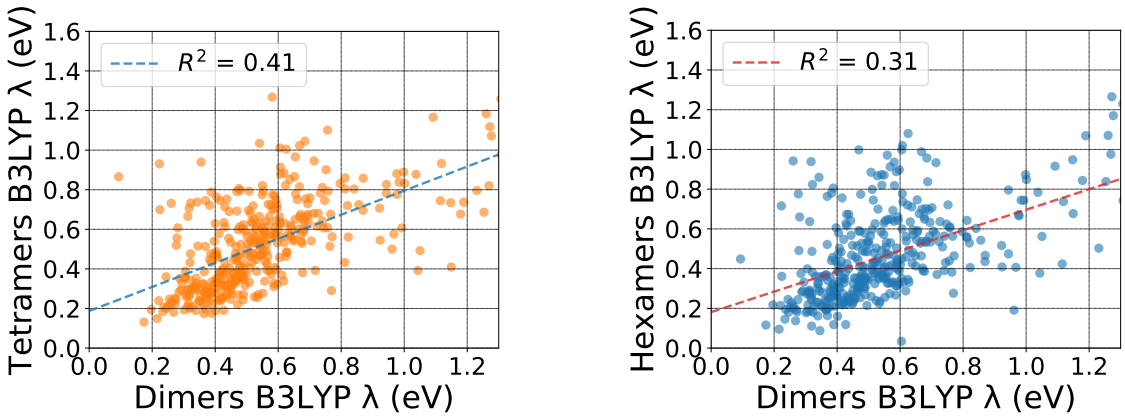


Figure A1: Correlation of B3LYP calculated λ between **(a)** dimers and tetramers, and **(b)** dimers and hexamers. Trendlines indicated robust linear regression fit.

(a) (b)

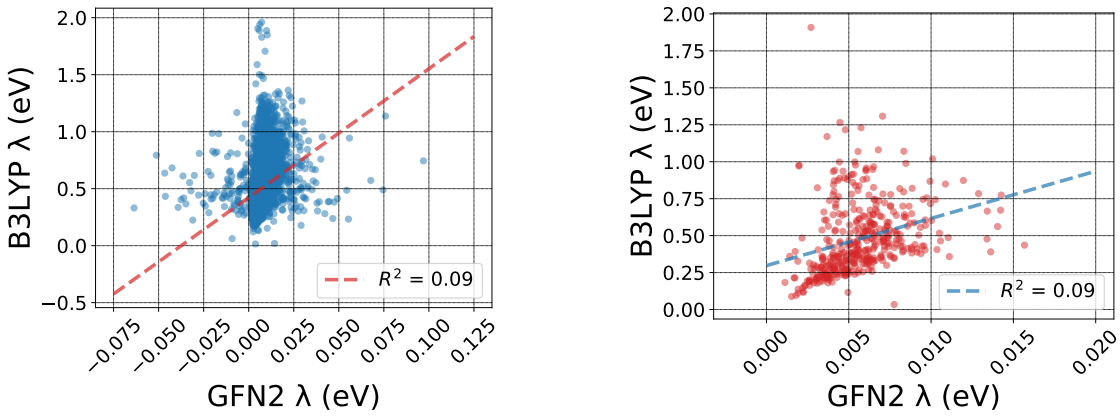


Figure A2: Correlation between λ calculated using B3LYP vs. λ calculated using GFN2 for **(a)** tetramers and **(b)** hexamers. Trendlines indicated robust linear regression fit.

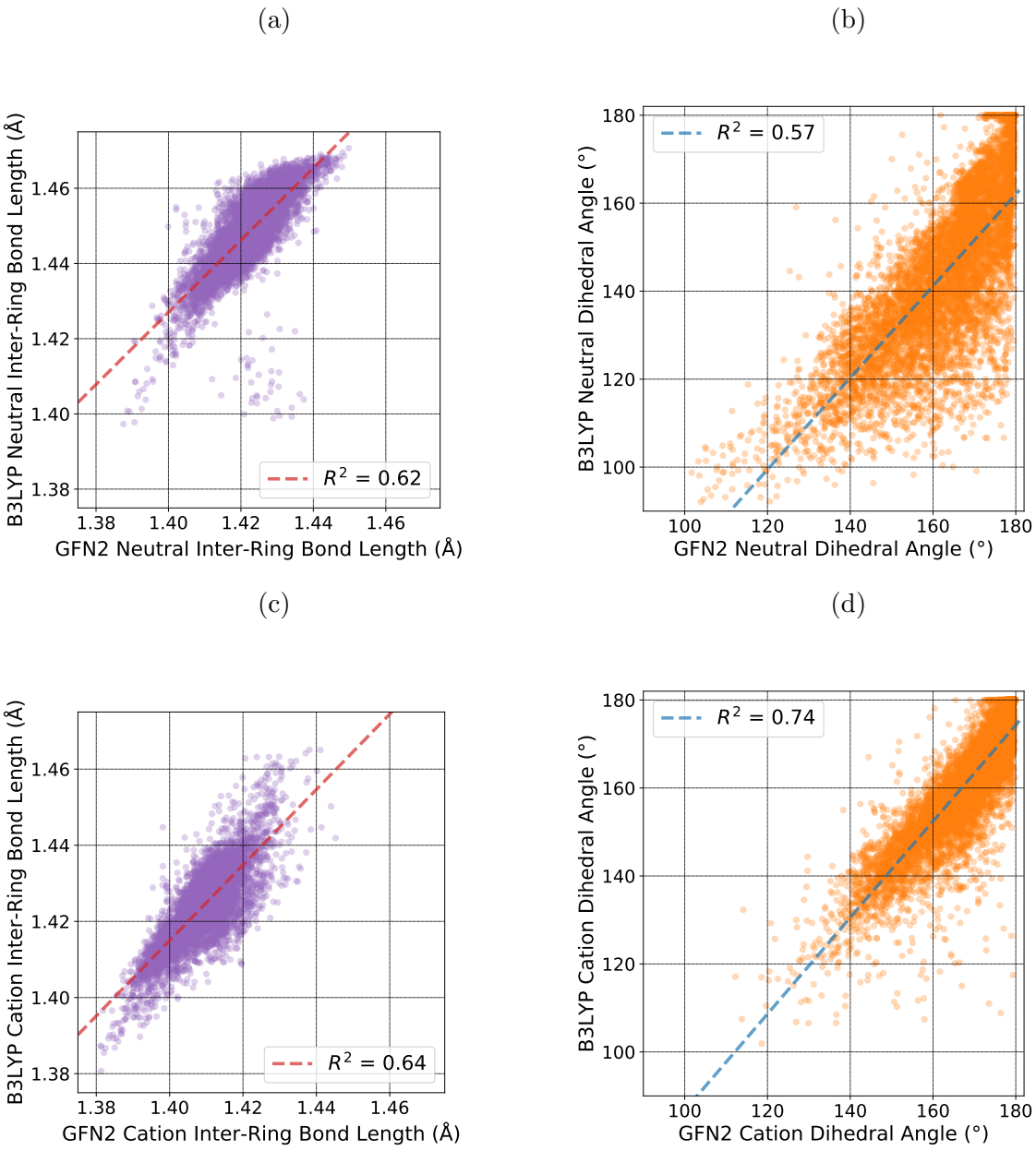


Figure A3: Correlation between the dihedral angle (**b, d**) and the inter-ring bond length (**a, c**) between the monomers calculated using B3LYP vs. GFN2 for the neutral (**a, b**) and cation (**c, d**) species. Trendlines indicated robust linear regression fit.

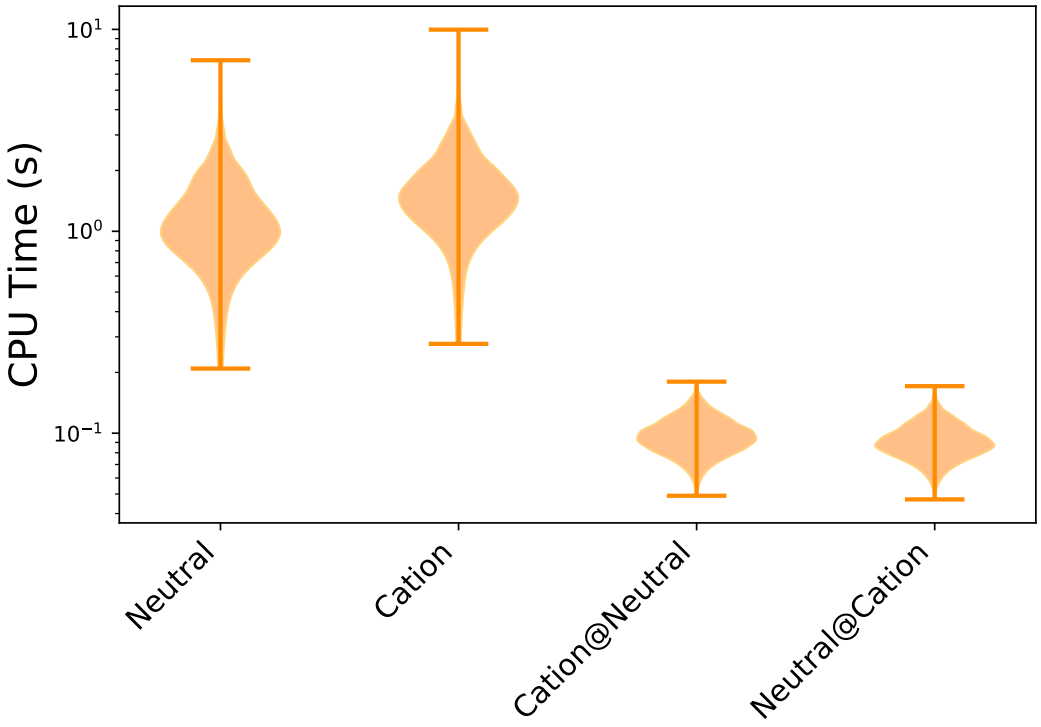


Figure A4: Calculation run time of the 4 different calculation for the dimers using GFN2. Note the logarithmic y-axis.

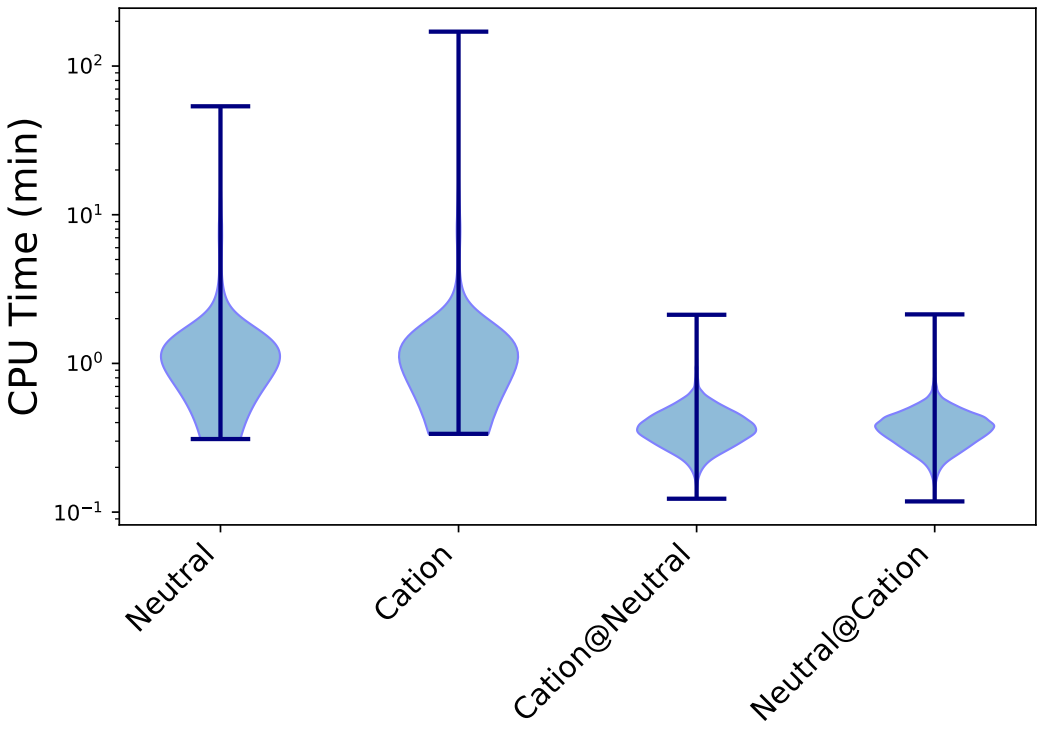


Figure A5: Calculation run time of the 4 different calculation for the tetramers using GFN2. Note the logarithmic y-axis.

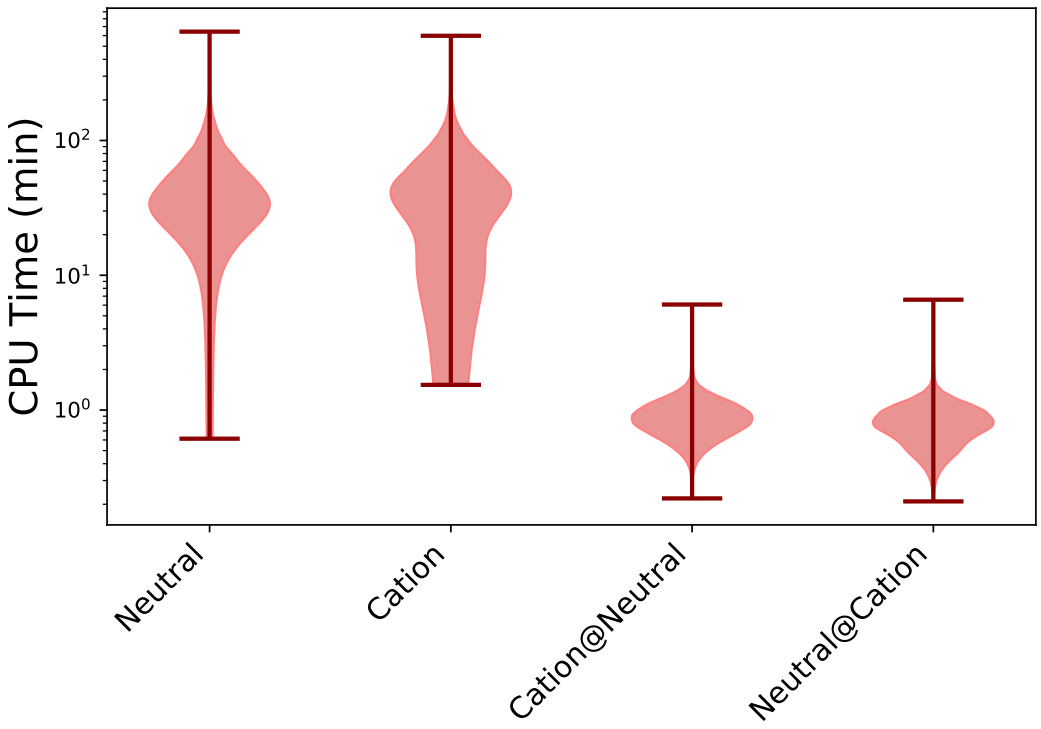


Figure A6: Calculation run time of the 4 different calculation for the hexamers using GFN2. Note the logarithmic y-axis.

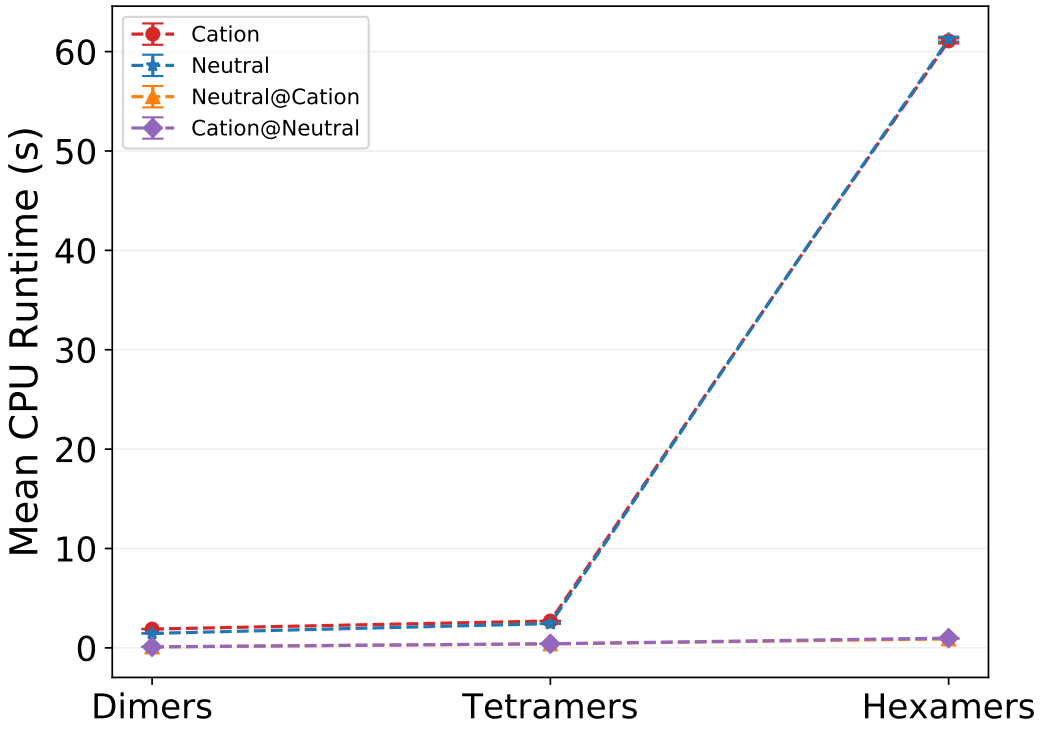


Figure A7: Mean run time for each of the 4 calculations for the dimers, tetramers, and hexamers using GFN2.

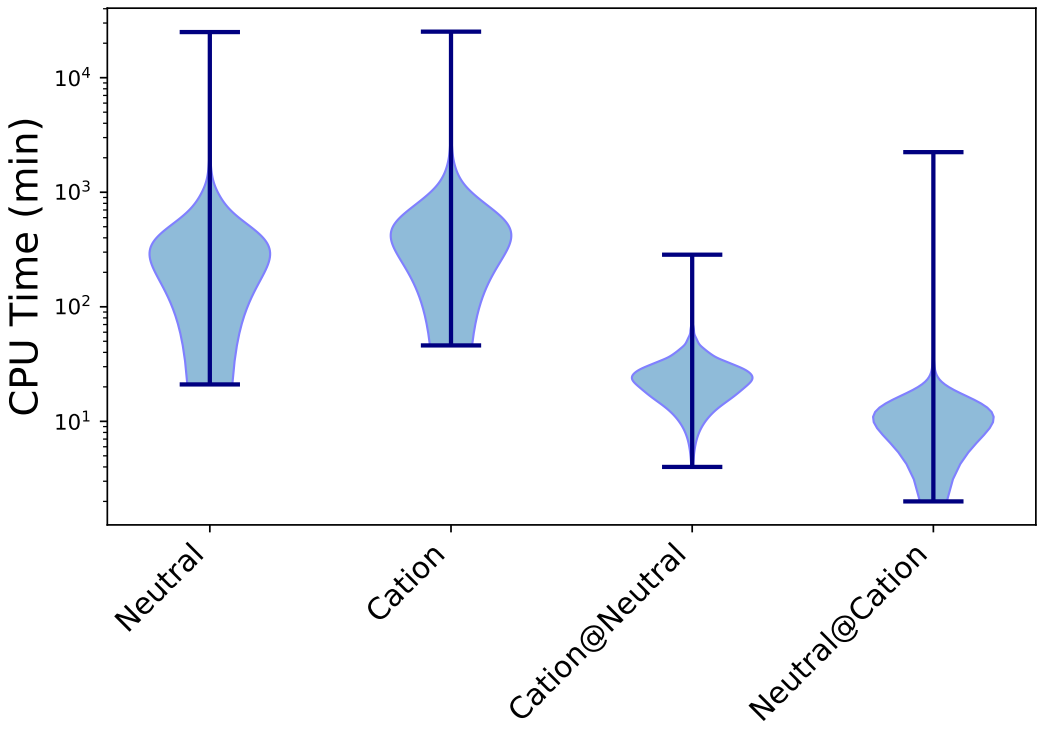


Figure A8: Calculation run time of the 4 different calculation for the tetramers using B3LYP. Note the logarithmic y-axis.

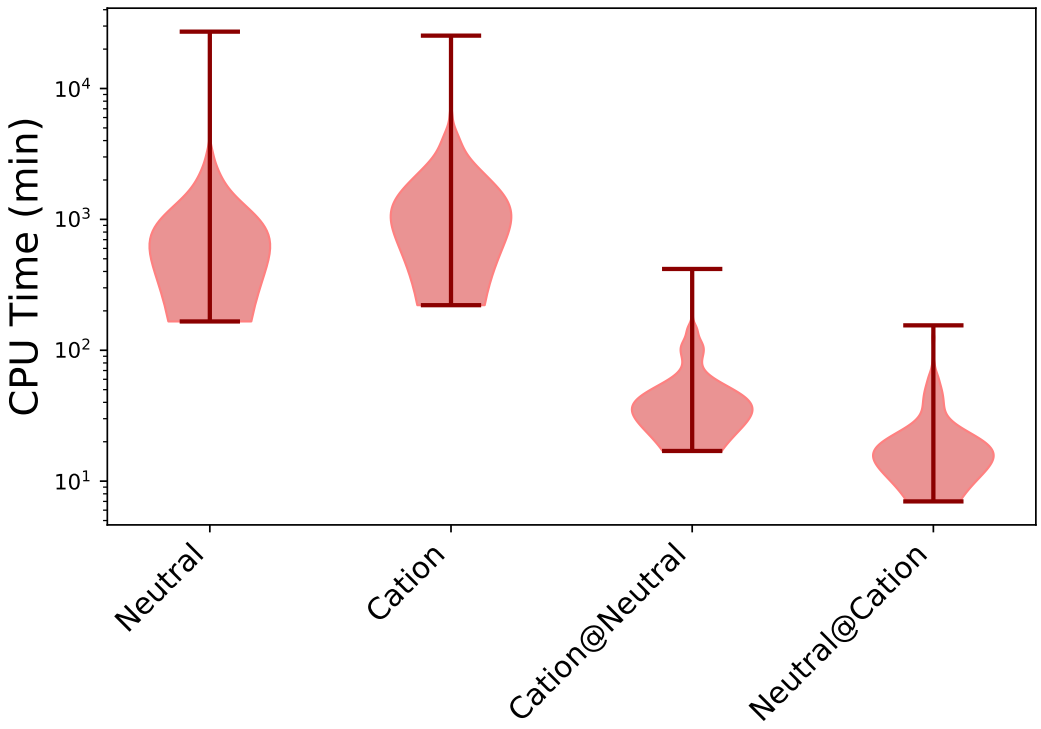


Figure A9: Calculation run time of the 4 different calculation for the hexamers using B3LYP. Note the logarithmic y-axis.

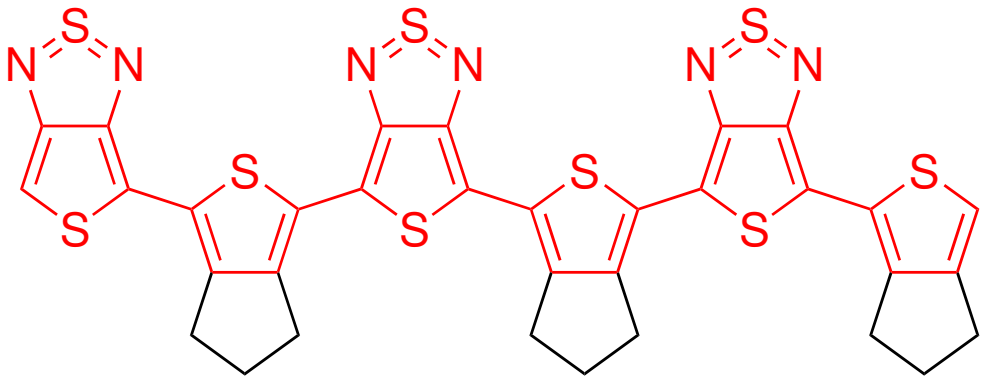


Figure A10: Example for the PiSystemSize feature, which counts the number of atoms in the longest continuous conjugated π-system. In this example of the hexamer of monomers 31 (cyclopentathiophene) and 47 (thiadiazolthiophene) - the 39 highlighted atoms in red are counted.

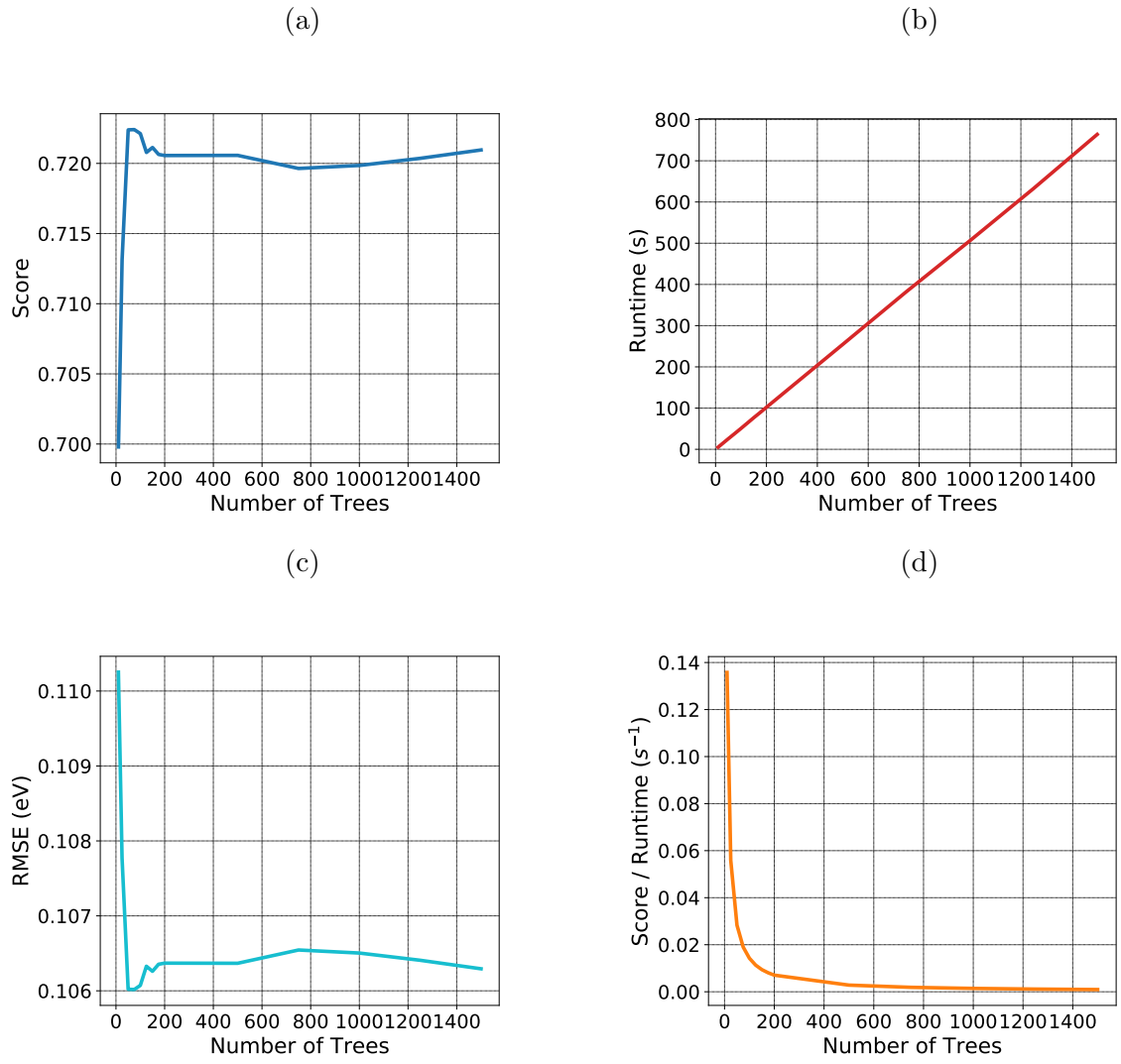


Figure A11: Random Forrest regression optimization: **(a)** score vs. number of trees, **(b)** run time vs. number of trees, **(c)** RMSE vs. number of trees and **(d)** score/run time vs. number of trees.

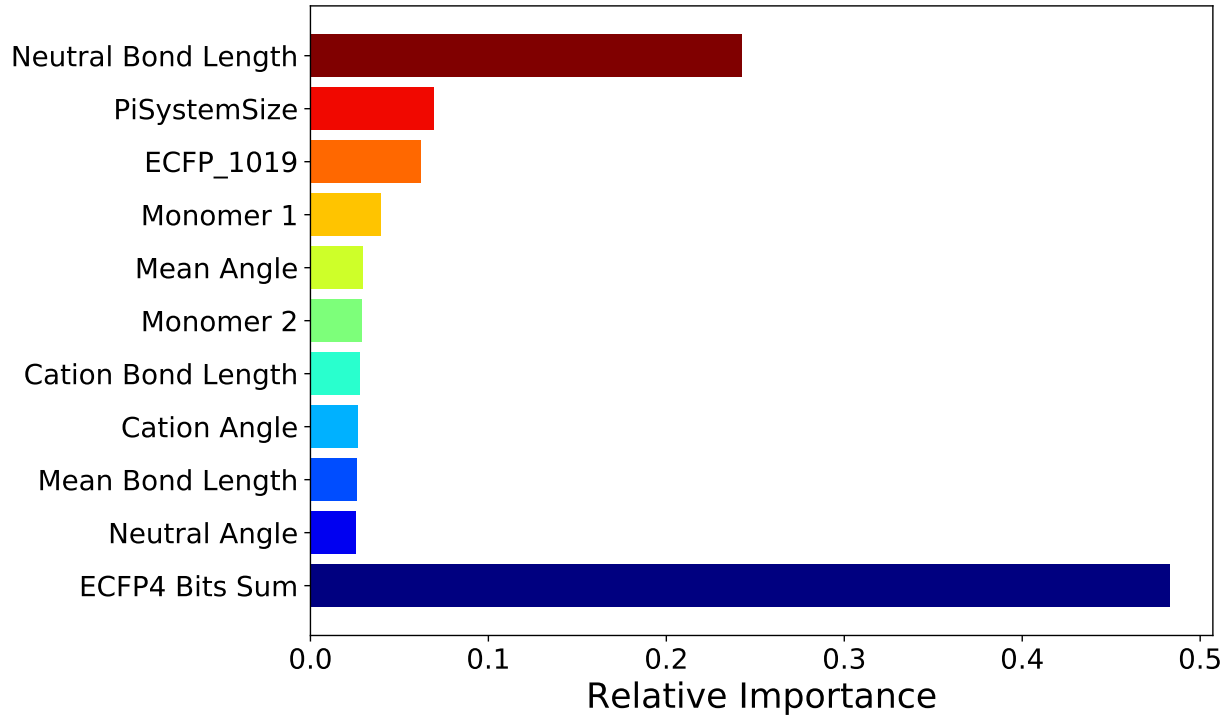


Figure A12: Relative feature importance of the top 10 features in the random forest model and the cumulative sum importance of all the ECFP4 bits.

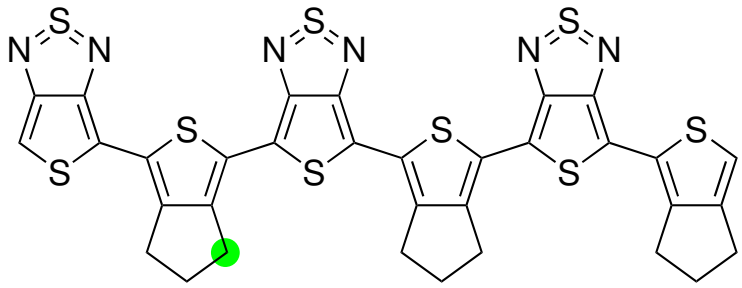


Figure A13: Example of the ECFP bit number 1019 which indicates the existence of an sp^3 hybridized carbon in the oligomer.

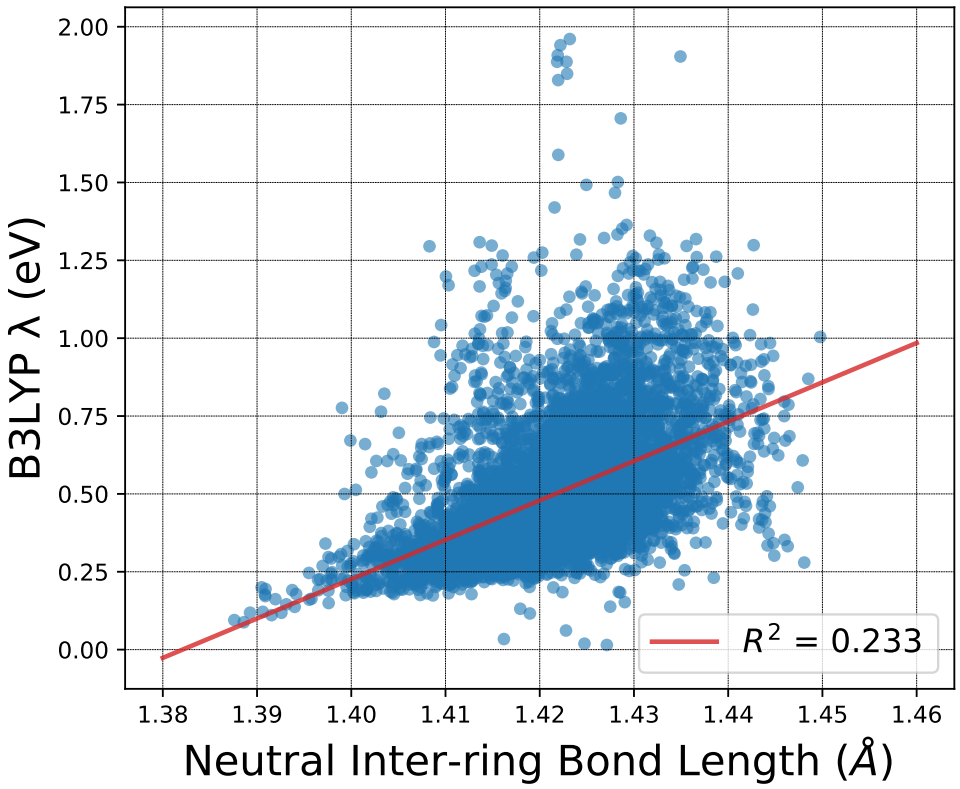
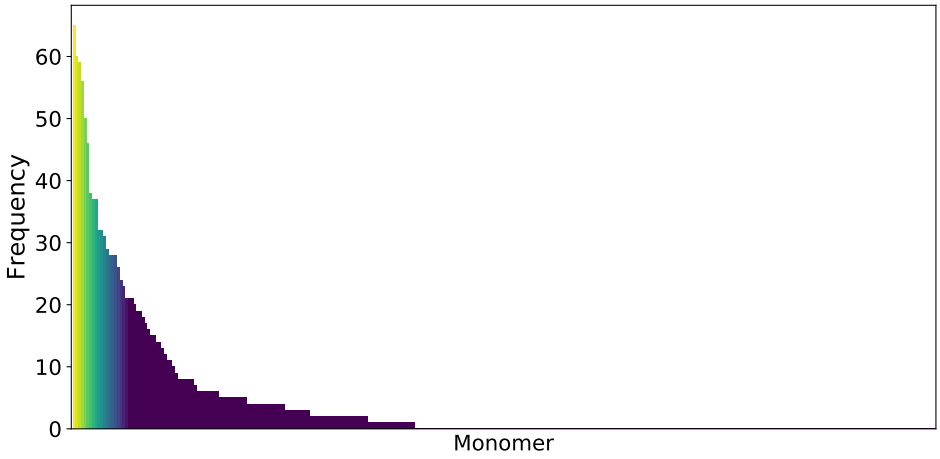


Figure A14: Correlation between the average neutral inter-ring bond length of the oligomers versus the B3LYP calculated λ.

(a)



(b)

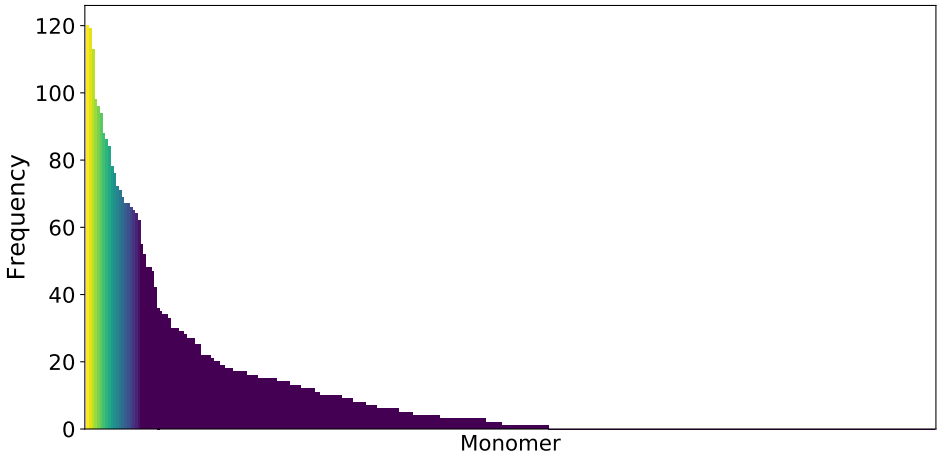


Figure A15: Histogram of monomers, sorted by frequency, for (a) tetramers and (b) hexamers with $\lambda < 0.3$ eV illustrating that only a small number of monomers are found frequently (compare to sorting by arbitrary monomer number in **(a)** 2.5c, and **(b)** 2.5d).

Table A1: The monomer numbers, the predicted and calculated B3LYP λ, the dihedral angles of the neutral and cation species, and the inter-ring bond length of both neutral and cation species for the 5 hexamers with the lowest B3LYP λ.

Monomer 1	Monomer 2	GFN2 Neutral Dihedral Angle (°)	GFN2 Cation Dihedral Angle (°)	B3LYP Neutral Dihedral Angle (°)	B3LYP Cation Dihedral Angle (°)	GFN2 Neutral Bond Length (Å)	GFN2 Cation Bond Length (Å)	B3LYP Neutral Bond Length (Å)	B3LYP Cation Bond Length (Å)
47	47	179.308	179.371	179.999	179.994	1.381	1.378	1.379	1.374
47	116	179.693	179.727	178.576	178.674	1.390	1.385	1.397	1.389
47	156	177.520	177.935	179.979	179.997	1.388	1.383	1.397	1.387
47	247	179.201	178.849	178.099	179.726	1.388	1.383	1.401	1.390
47	217	179.460	179.493	175.533	179.994	1.393	1.387	1.403	1.391

Appendix B Strategies for Computer-Aided Discovery of Novel Open-Shell Polymers

B.1 Code and Data Availability

Full code, data files, and analysis notebooks are available at `https://github.com/Shualdon/GST`

B.2 Supplementary Information

$$\Delta E_{T-S} = 1.27 \times (HOMO - LUMO\ Gap_{singlet}) - 4.88 \tag{5}$$

EQ. S5: The best fit linear correlation equation between ΔE_{T-S} and $HOMO - LUMO\ Gap_{singlet}$. All values are in eV.

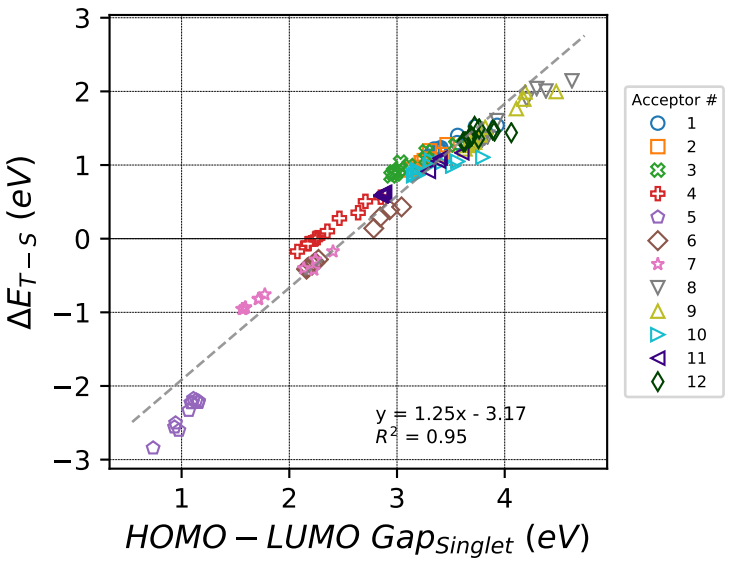
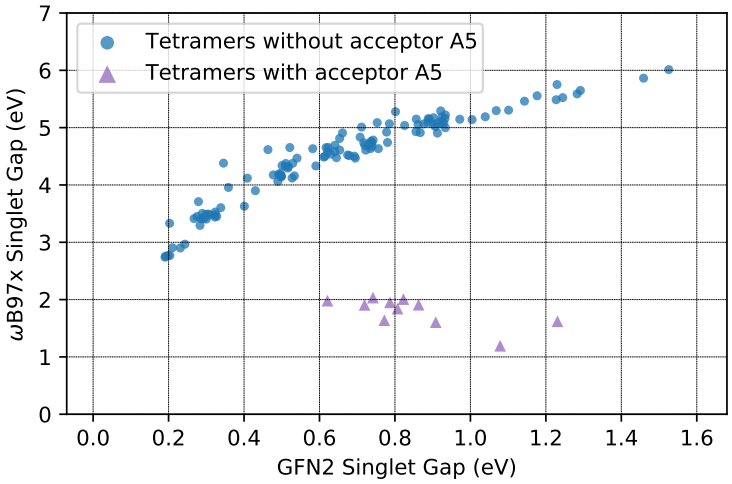


Figure B1: Correlation plots between the difference of the Triplet and Singlet energies of each oligomer versus its the HOMO-LUMO gap of the singlet species, both in eV, calculated using the CAM-B3LYP functional, grouped by the acceptor number. Linear best-fit line is shown as a dashed gray line.

(a)



(b)

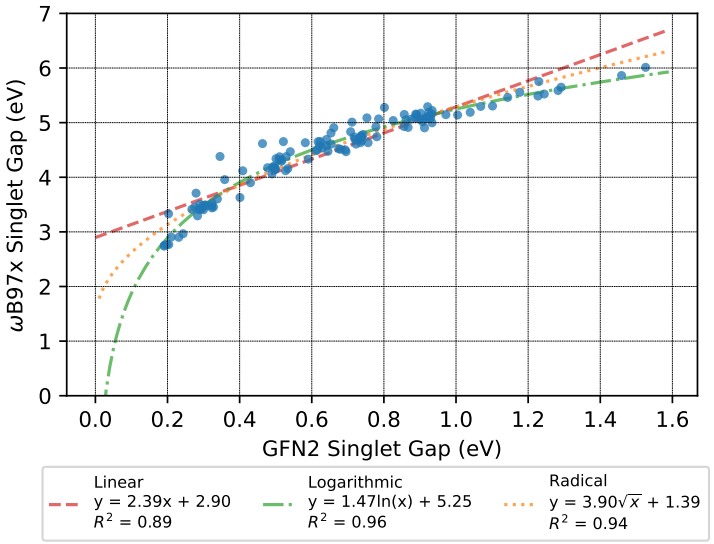
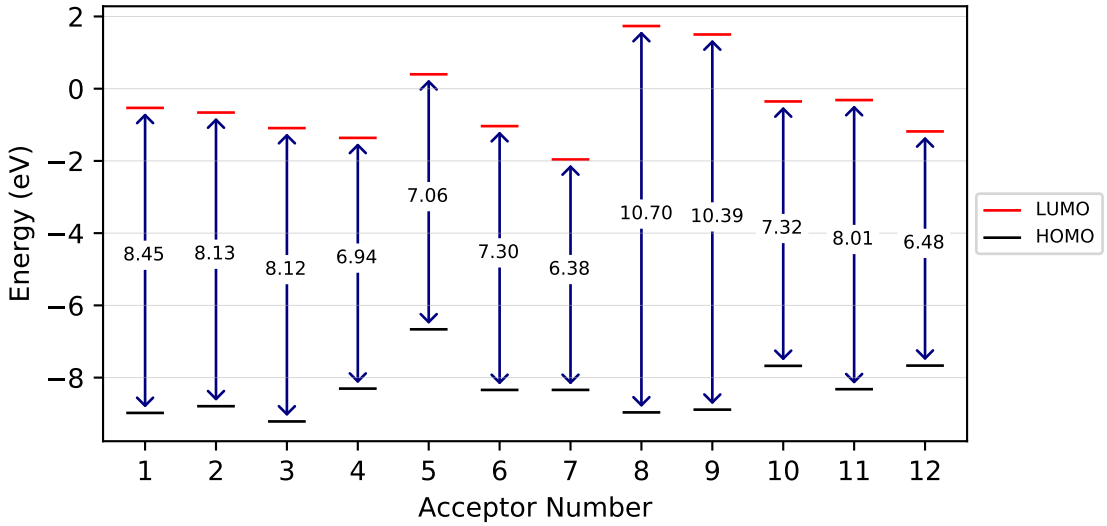


Figure B2: Correlation between the singlet HOMO-LUMO gap calculated using ωB97X-D versus GFN2-xTB, (**a**) showing all tetramers, where tetramers that contain acceptors A5 are shown in purple triangles, (**b**) showing only tetramers that do not contain acceptor A5. Linear, logarithmic, and radical functions were fit to the data in order to find the highest correlated function.

(a)



(b)

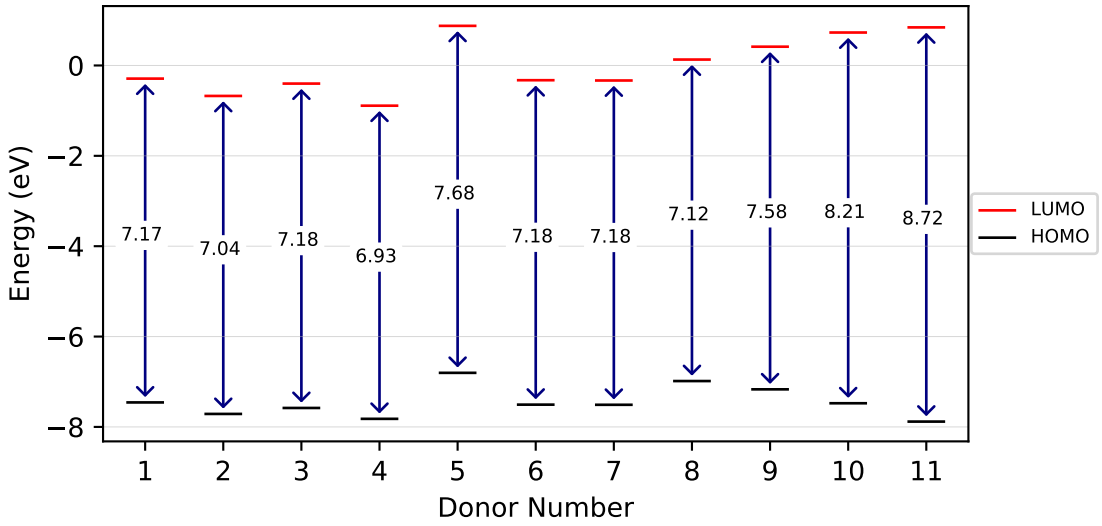


Figure B3: The HOMO and LUMO energies, in eV, of the (**a**) acceptor monomers and (**b**) donor monomers, with the HOMO-LUMO gap energy, also in eV, in the center of each graph.

Table B1: The slope of the linear best-fit function and its coefficient of determination (R^2) between the inter-monomer bond length and ΔE_{T-S} for tetramers that share acceptors.

Acceptor Number	Best-Fit Slope (eV/Å)	R^2
1	17.81	0.37
2	41.92	0.92
3	38.90	0.89
4	48.16	0.80
5	-41.93	0.33
6	3.14	0.02
7	2.9	0.02
8	-3.33	0.01
9	90.72	0.69
10	18.15	0.52
11	56.13	0.88
12	-2.43	0.20

Appendix C Using Genetic Algorithms to Discover Novel Ground-State Triplet Conjugated Polymers

C.1 Code and Data Availability

Full code, data files, and analysis notebooks are available at `https://github.com/Shualdon/GST_GA`

C.2 Supplementary Information

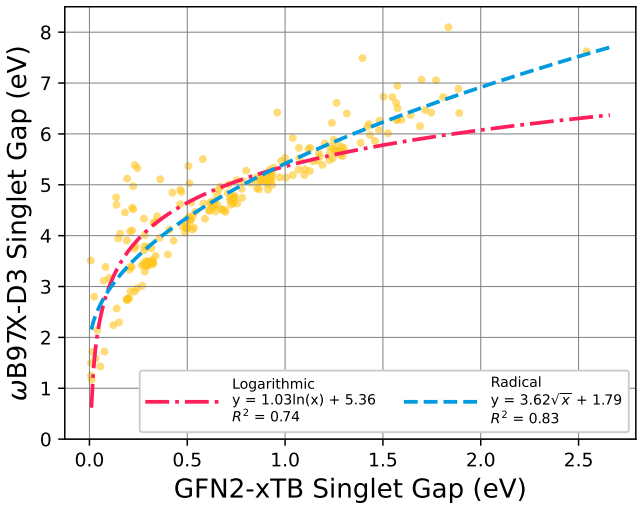


Figure C1: Correlation between HOMO-LUMO gaps calculated using GFN2-xTB versus ωB97X-D3. The logarithmic, in red dash-dotted line, and the radical, in blue dashed line, best-fit functions, with their respective equations and coefficient of determination (R^2), are shown.

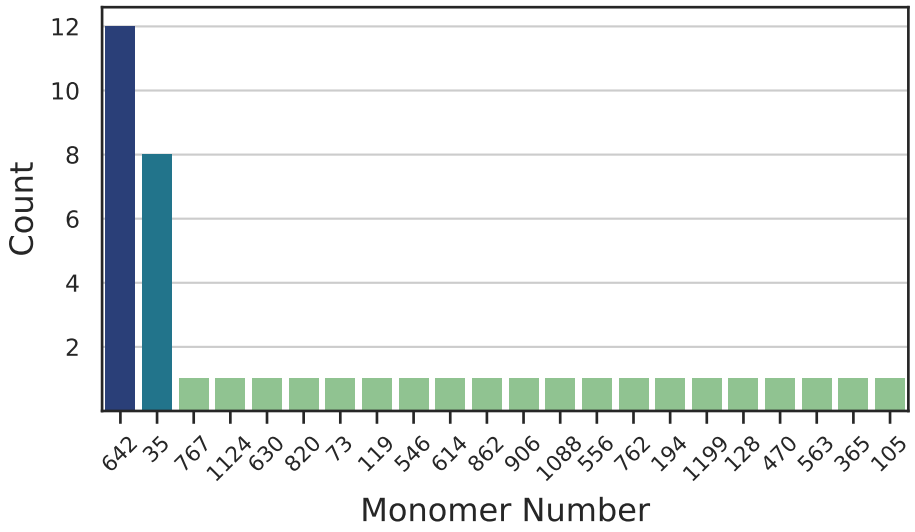


Figure C2: The number of times a monomer was part of the top 20 oligomers, i.e. with the lowest xTB-GNF2 HOMO-LUMO gap, found in all 10 GA runs. Each of the oligomers had either monomer 642 or monomer 35 as one of their monomers.

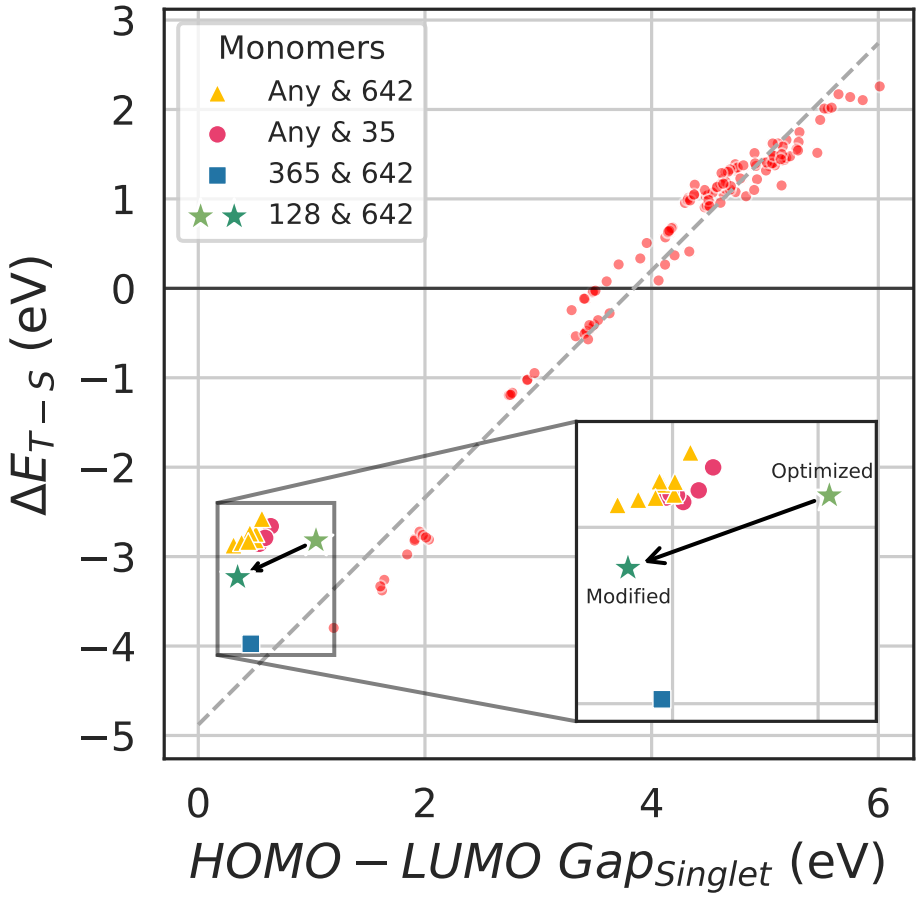
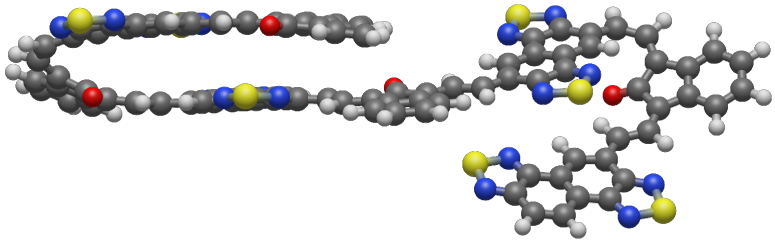


Figure C3: Correlation between the singlet HOMO-LUMO gap and ΔE_{T-S} of the 16 out of the top 20 oligomers. The oligomers are grouped by the common monomers—35 (in yellow triangles) and 624 (in pink circles). The outlier of monomer 128 and 642 is indicated with light green star with the values of its optimized geometry (at the arrow's tail) and with a dark green star at its values in the modified geometry (at the arrow's head). The outlier of monomers 365 and 642 is indicated in a blue square. A zoomed-in inset of the relevant part is shown. The red points are the data points from the previous study, with the best fit line for those points shown in dashed gray line.

(a)



(b)

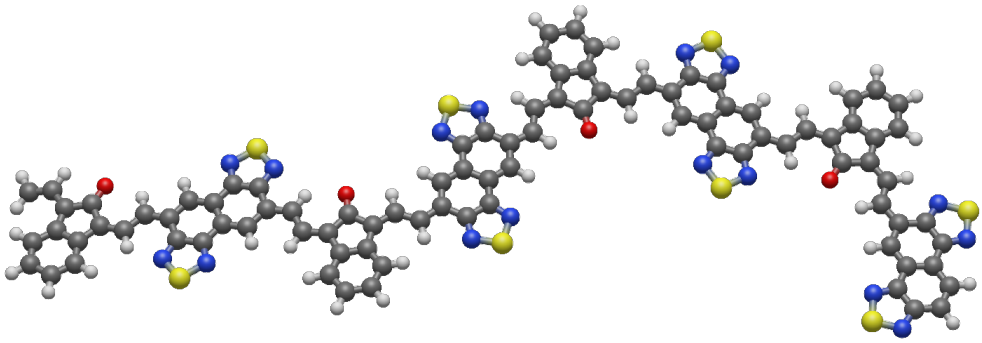


Figure C4: **a** A side view of the optimized, folded conformation of oligomer 128_642, **b** A top view of the modified, flat conformation of oligomer 128_642.

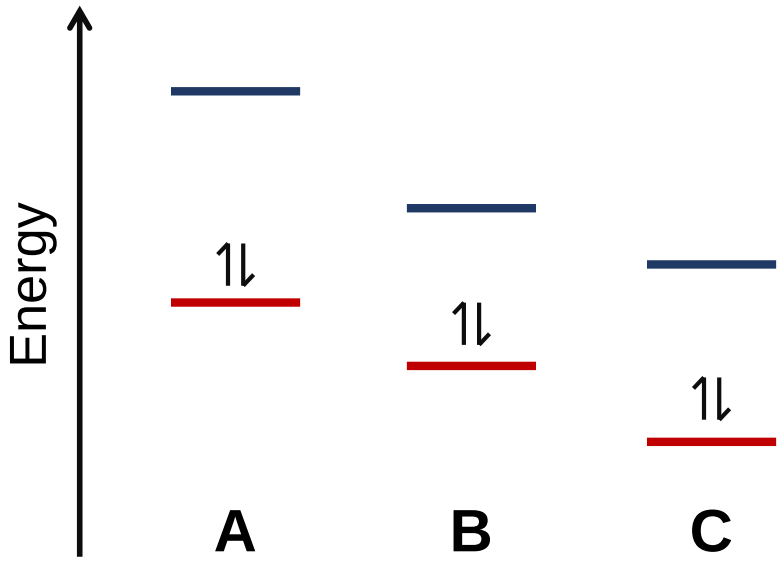
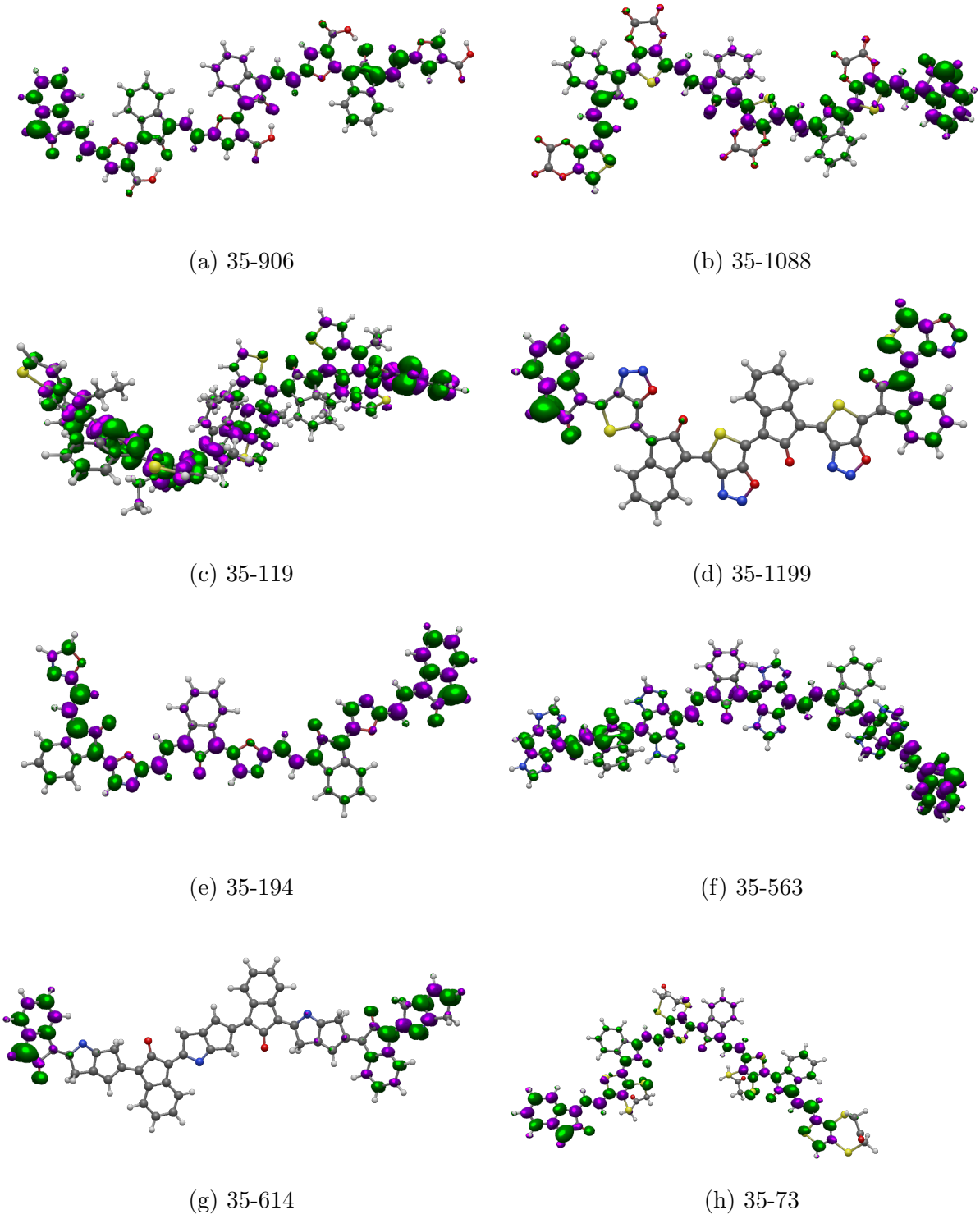


Figure C5: Relative HOMO (in red) and LUMO (in blue) levels of hypothetical monomers A, B, and C. If monomers A and B were to combine in an polymer monomer A will be the donor while monomer B will be the acceptor, as the HOMO level of monomer A is relatively higher in energy than monomer B. However, if monomers B and C were to combine to make a polymer then monomer B will be the donor while monomer C will be the acceptor. Monomer B can behave as either a donor or acceptor, depending on which monomer it is paired with.



(a) 35-906

(b) 35-1088

(c) 35-119

(d) 35-1199

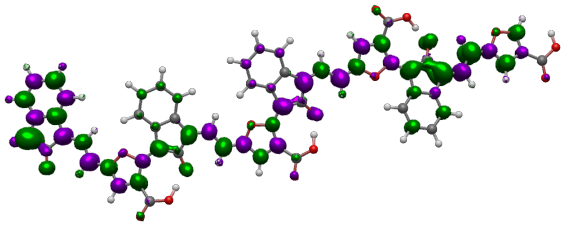
(e) 35-194

(f) 35-563

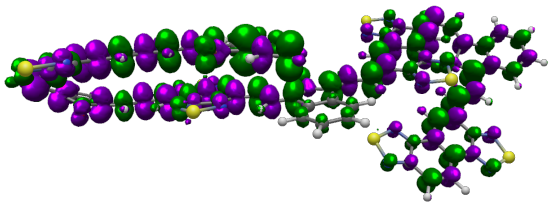
(g) 35-614

(h) 35-73

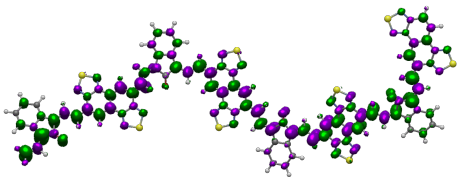
Figure C6: Spin density plots of the top 20 oligomers. Isosurface value is 0.002 a.u.



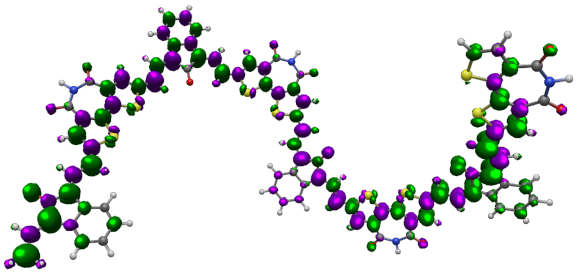
(i) 35-906



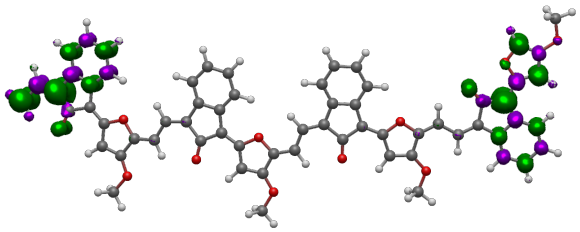
(j) 642-128



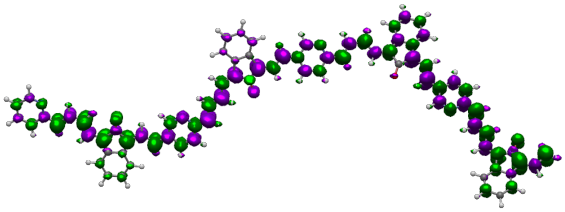
(k) 642-128 modified



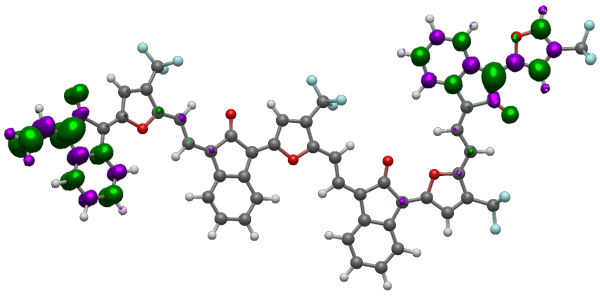
(l) 642-365



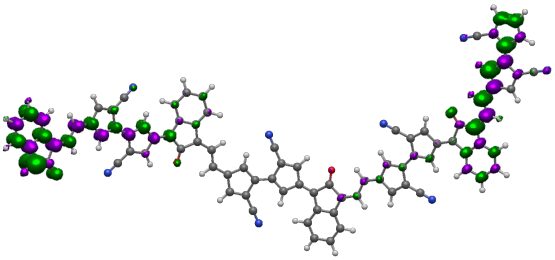
(m) 642-470



(n) 642-546

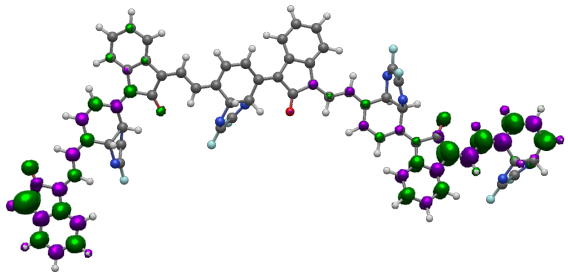


(o) 642-556

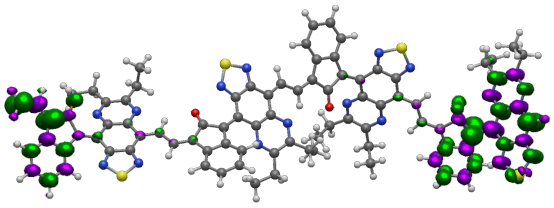


(p) 642-105

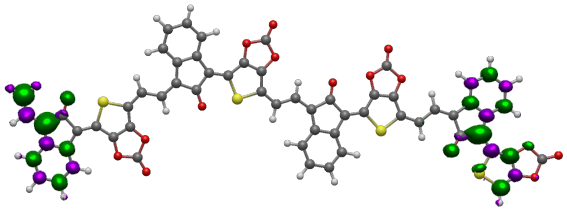
Figure C6: Cont. Spin density plots of the top 20 oligomers. Isosurface value is 0.002 a.u.



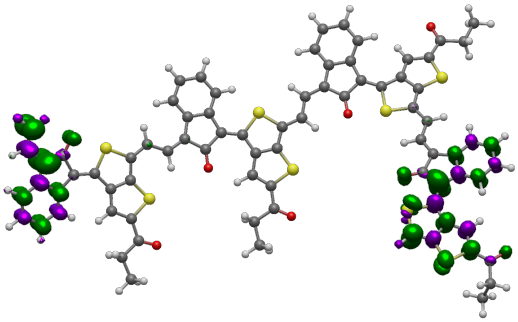
(q) 642-1124



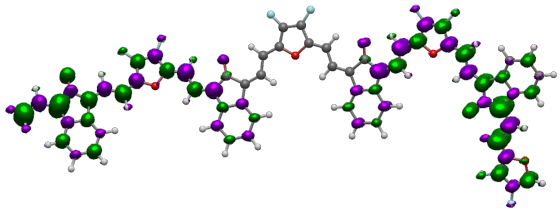
(r) 642-762



(s) 642-767



(t) 642-820



(u) 642-862

Figure C6: Cont. Spin density plots of the top 20 oligomers. Isosurface value is 0.002 a.u.

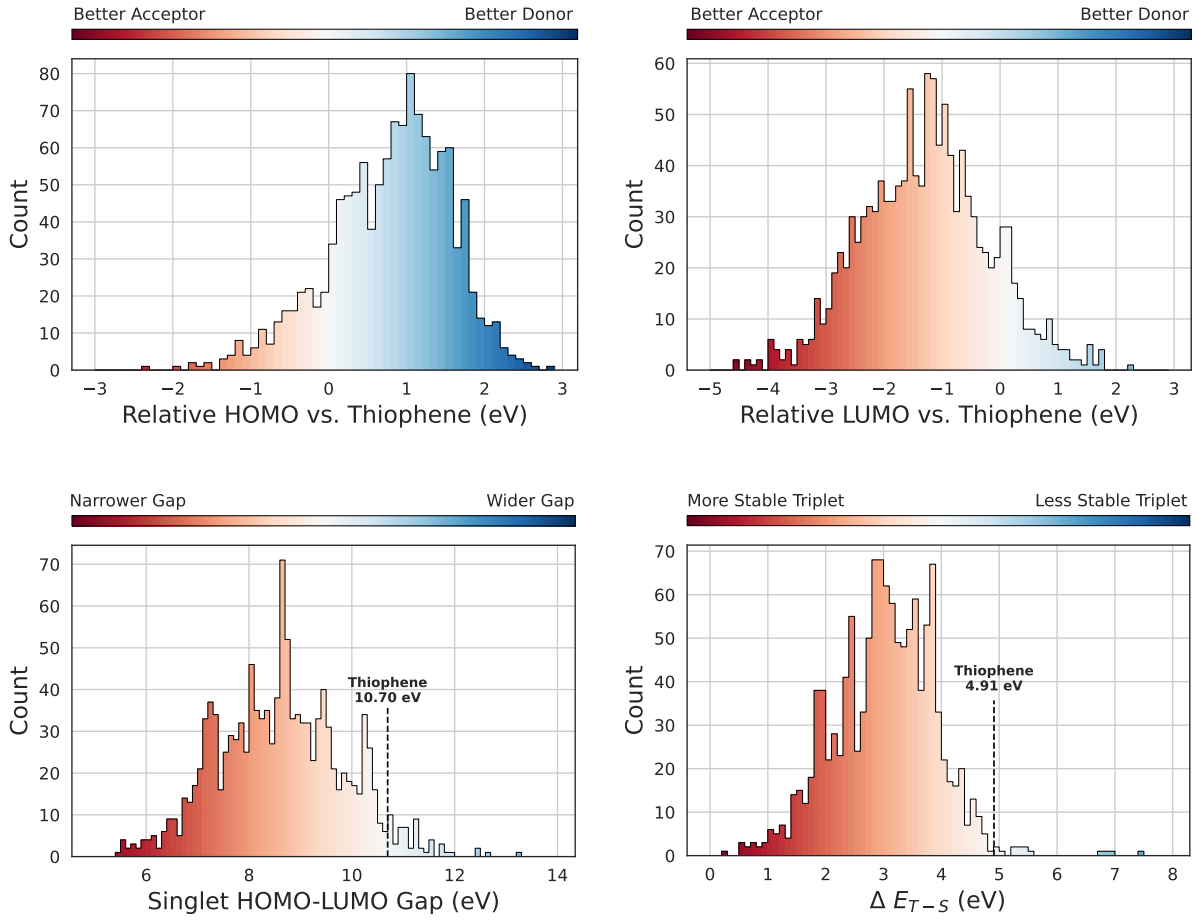


Figure C7: Top Left: a histogram of the monomers' HOMO eigenvalue relative to thiophene's HOMO eigenvalue. Top Right: a histogram of the monomers' LUMO eigenvalue relative to thiophene's LUMO eigenvalue. Bottom Left: A histogram of the monomers' HOMO-LUMO gap. Thiophene's HOMO-LUMO gap is marked for reference. Bottom Right: A histogram of the monomers' electronic energy difference between the triplet and singlet ground states. A Lower value correlates to a more stable triplet ground-state. Thiophene's ΔE_{T-S} is marked for reference.

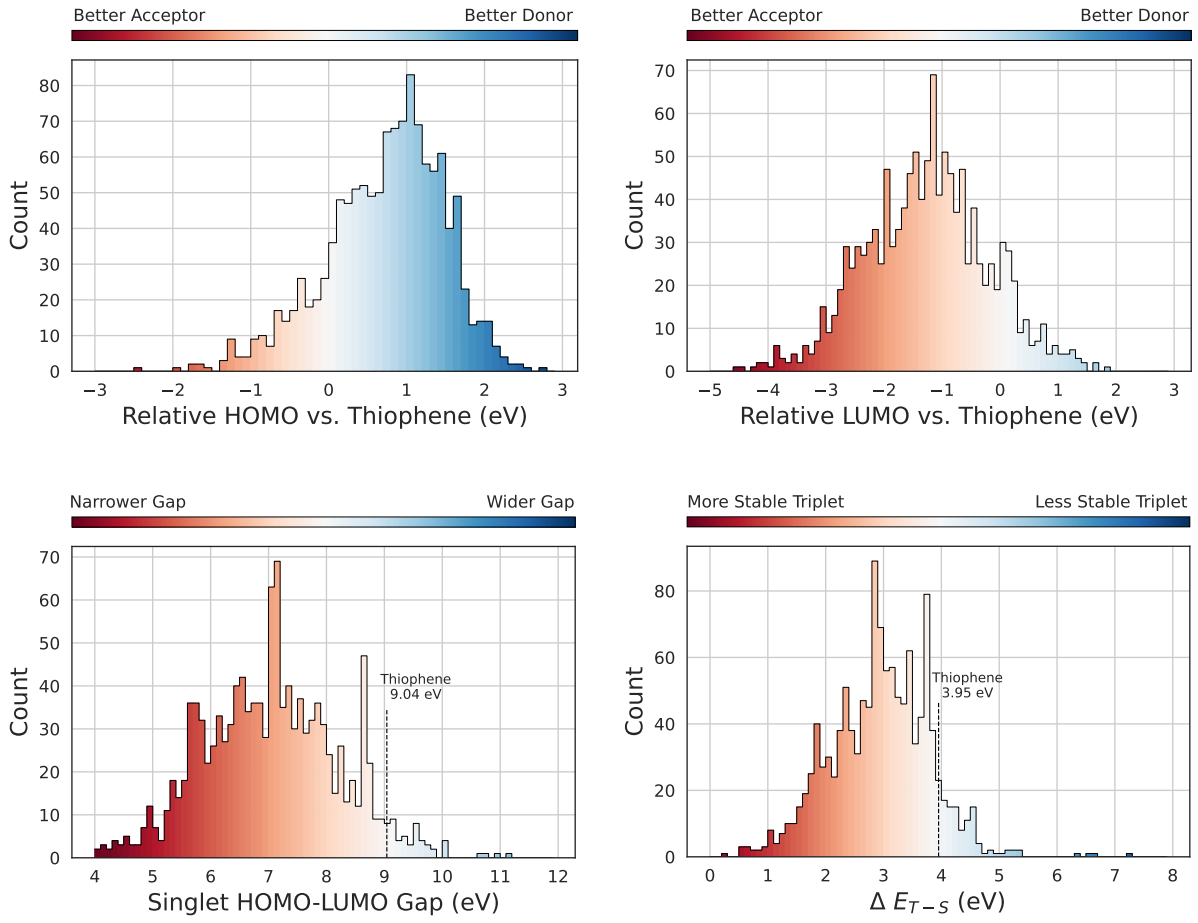


Figure C8: CAM-B3LYP single point calculations on the monomers that show similar distributions to the ωB97X-D3 single point calculations in Figure C7. Top Left: a histogram of the monomers' HOMO eigenvalue relative to thiophene's HOMO eigenvalue. Top Right: a histogram of the monomers' LUMO eigenvalue relative to thiophene's LUMO eigenvalue. Bottom Left: A histogram of the monomers' HOMO-LUMO gap. Thiophene's HOMO-LUMO gap is marked for reference. Bottom Right: A histogram of the monomers' electronic energy difference between the triplet and singlet ground states. A Lower value correlates to a more stable triplet ground-state. Thiophene's $\Delta\ E_{T-S}$ is marked for reference.

Table C1: The HOMO level (relative to thiophene's), LUMO level (relative to thiophene's), the HOMO-LUMO gap and the ΔE_{T-S} of the 10 most common monomers from all the GA runs (Figure 4.3).

Monomer	Relative HOMO	Relative LUMO	HOMO-LUMO Gap	ΔE_{T-S}
35	1.05	-3.35	6.31	0.71
77	1.71	-2.54	6.45	0.92
115	-0.87	-4.58	6.99	1.63
187	-0.19	-4.51	6.38	1.28
221	0.15	-3.35	7.20	1.13
642	1.56	-3.49	5.65	0.28
686	1.33	-2.51	6.86	1.60
778	2.65	-2.50	5.55	0.68
1029	-0.66	-3.62	7.74	3.41
1212	0.62	-3.69	6.38	1.17

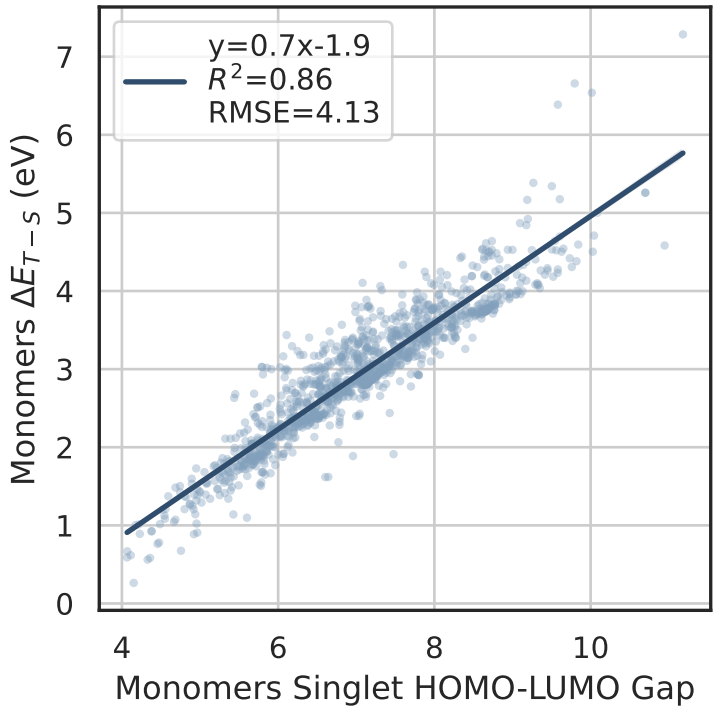


Figure C9: Correlation between the monomers' singlet HOMO-LUMO gap and the stability of their triplet ground state, ΔE_{T-S}.

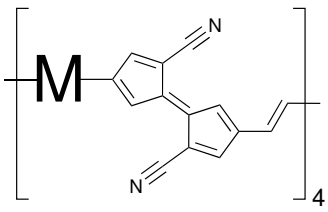


Figure C10: Visualization of the oligomers constructed by some monomer, **M**, and monomer number 630, for Figure 4.5.

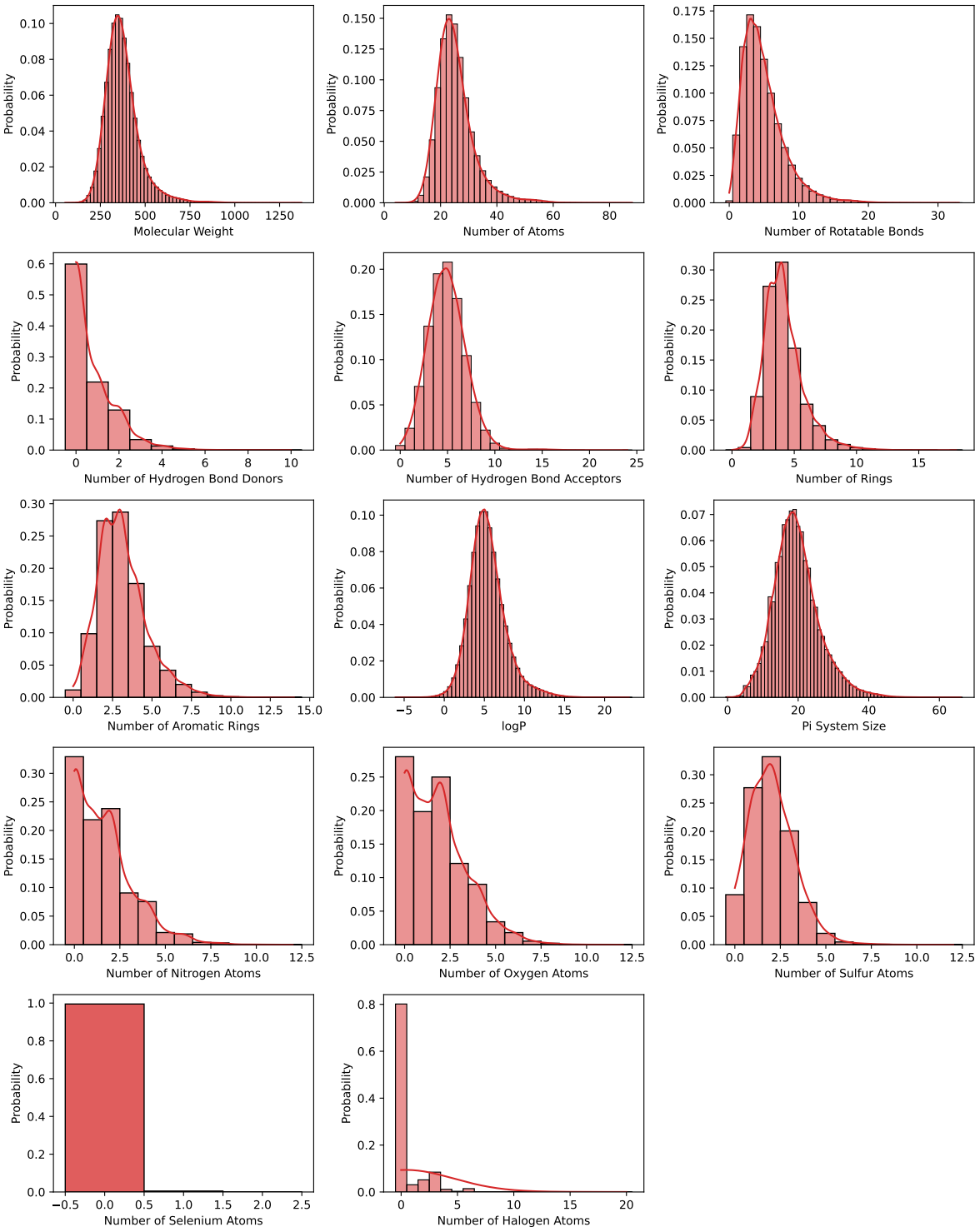


Figure C11: Various descriptors of all (∼1.5 million) possible monomer pairs.

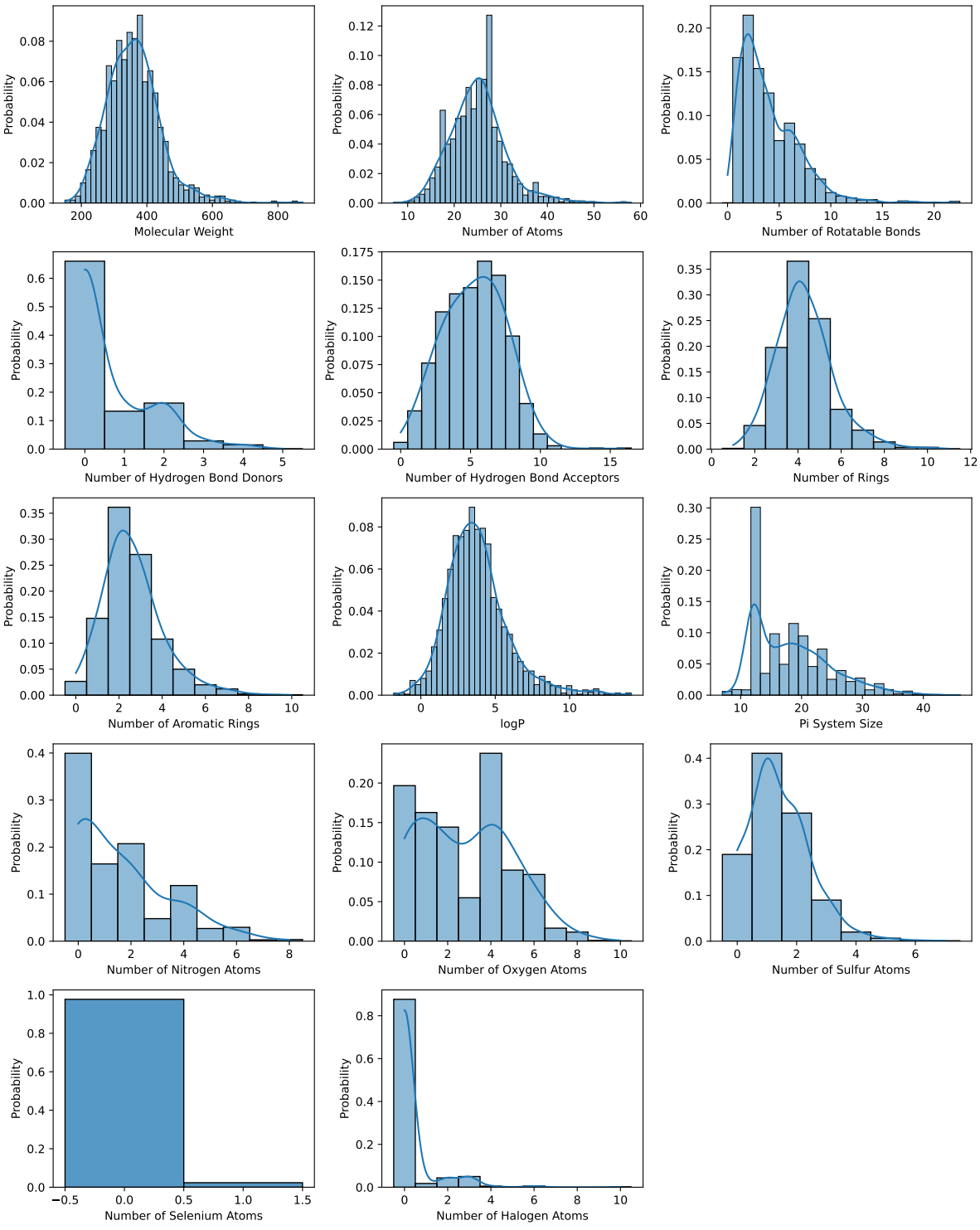


Figure C12: Various descriptors of the monomers pairs of the oligomers with GFN2-xTB–calculated HOMO-LUMO gap smaller than 0.2 eV that were generated in any of the GA runs.

Table C2: Full data for Figure 4.5. Monomers number that were combined with 630 to create an oligomer, each monomer's relative HOMO, relative LUMO, its singlet HOMO-LUMO gap and its ΔE_{T-S}, and the ΔE_{T-S} of the full oligomer. The table is sorted by the ascending oligomer's ΔE_{T-S}.

Monomer Number	Monomer Relative HOMO	Monomer Relative LUMO	Monomer Singlet Gap	Monomer ΔE_{T-S}	Oligomer ΔE_{T-S}
261	0.73	-3.46	6.51	1.05	**-2.96**
224	0.06	-4.26	6.37	1.29	**-2.75**
1172	1.66	-3.14	5.89	0.80	**-2.72**
768	1.77	-2.75	6.17	1.08	**-2.70**
141	1.92	1.19	9.97	3.98	**-2.69**
1064	2.29	-2.28	6.13	1.56	**-2.42**
1181	2.18	-2.33	6.18	1.21	**-2.10**
200	2.34	-2.29	6.06	1.54	**-2.03**
775	1.91	-3.20	5.59	0.64	**-1.88**
710	1.21	-2.24	7.24	1.92	**-1.79**
663	0.63	-2.04	8.04	2.48	**-1.51**
40	0.93	-1.98	7.78	2.38	**-1.43**
1027	1.16	-3.00	6.54	0.96	**-1.06**
630	0.14	-4.13	6.43	1.32	**-1.00**
1128	-0.41	-4.33	6.78	1.68	**-1.00**
337	0.76	-1.47	8.46	3.44	**-0.99**
913	2.48	-2.32	5.90	1.01	**-0.93**
1002	0.18	-0.40	10.12	3.46	**-0.71**
105	-0.33	-4.11	6.92	1.56	**-0.70**
639	0.58	-2.75	7.37	2.16	**-0.46**
804	1.06	-0.81	8.82	2.99	**-0.33**
75	2.59	-2.00	6.10	1.04	**-0.30**
730	1.46	-3.22	6.02	1.15	**-0.15**
728	1.74	-1.78	7.17	1.81	**-0.14**
1204	2.83	0.68	8.54	3.21	**-0.08**
1067	1.55	-3.49	5.66	0.91	**-0.04**
722	1.52	-3.13	6.04	1.11	**-0.03**
203	2.49	-2.43	5.78	0.56	**0.01**
569	1.09	-3.18	6.43	1.43	**0.02**
395	1.73	-3.48	5.50	0.55	**0.04**
697	1.31	-3.32	6.06	1.41	**0.05**
225	2.22	0.23	8.70	3.15	**0.07**
526	1.62	0.83	9.91	3.89	**0.16**
466	2.41	-2.76	5.53	1.06	**0.19**
1054	0.26	-1.05	9.39	3.57	**0.20**
332	-0.52	-4.30	6.92	1.62	**0.21**
305	2.55	-2.35	5.81	0.58	**0.23**
210	2.37	-0.28	8.04	3.06	**0.23**
137	-0.08	-3.91	6.87	1.67	**0.32**
176	2.36	-2.76	5.58	1.07	**0.35**
588	1.29	-1.90	7.50	2.76	**0.35**
670	2.23	-1.06	7.41	2.36	**0.39**
418	2.24	-0.72	7.74	2.48	**0.40**
503	0.20	-3.99	6.51	1.67	**0.41**
34	1.93	-0.22	8.55	3.26	**0.46**
349	0.24	-3.97	6.49	1.57	**0.47**
954	2.25	1.01	9.45	3.85	**0.50**
343	2.18	-2.81	5.71	0.89	**0.51**

Table C2 (continued).

Monomer Number	Monomer Relative HOMO	Monomer Relative LUMO	Monomer Singlet Gap	Monomer ΔE_{T-S}	Oligomer ΔE_{T-S}
534	-0.22	-1.91	9.01	3.13	**0.53**
100	0.48	-0.78	9.44	3.57	**0.56**
104	2.22	-2.77	5.71	0.88	**0.58**
618	0.04	-2.01	8.65	2.95	**0.62**
389	-0.97	-2.11	9.56	3.65	**0.63**
457	-1.14	-1.62	10.22	3.85	**0.63**
869	-1.34	-1.93	10.11	3.80	**0.65**
1	1.45	0.65	9.90	3.90	**0.65**
723	-1.28	-1.88	10.10	5.27	**0.67**
591	-1.19	-1.26	10.63	4.07	**0.69**
1203	-1.35	-2.08	9.96	3.82	**0.71**
764	0.72	-2.57	7.40	2.34	**0.71**
472	-1.59	-2.57	9.72	3.82	**0.72**
0	-1.23	-1.18	10.74	4.13	**0.76**
160	0.14	-3.80	6.76	1.50	**0.76**
976	-1.64	-2.61	9.72	4.24	**0.79**
5	0.04	-3.76	6.89	1.88	**0.80**
41	0.05	-1.51	9.14	1.95	**0.82**
1071	1.40	-1.29	8.00	3.28	**0.87**
242	0.85	-1.01	8.84	4.20	**0.88**
515	-1.37	-2.58	9.48	3.75	**0.89**
1030	-1.20	-2.47	9.42	3.58	**0.90**
826	1.34	-1.29	8.06	2.46	**0.90**
519	-1.27	-0.13	11.84	5.09	**0.91**
413	0.96	1.52	11.25	4.62	**0.93**
966	-1.10	-3.79	8.00	3.25	**0.95**
471	1.74	-1.79	7.16	1.82	**0.96**
1136	-1.71	-1.27	11.13	4.36	**0.97**
893	-0.99	-3.82	7.86	3.27	**0.98**
746	0.72	1.14	11.12	4.37	**0.99**
346	-0.77	-3.85	7.62	3.00	**0.99**
783	0.36	-0.51	9.83	3.59	**1.03**
1135	-1.21	-1.17	10.74	4.39	**1.04**
996	1.48	-2.22	6.99	1.81	**1.06**
1095	0.47	-2.04	8.18	2.80	**1.06**
943	0.64	-2.74	7.32	2.10	**1.07**
67	1.65	0.58	9.64	3.85	**1.08**
725	-2.39	-3.87	9.22	3.61	**1.08**
533	0.27	0.84	11.27	4.31	**1.08**
581	-1.55	-3.94	8.30	3.32	**1.09**
504	1.00	1.38	11.07	4.11	**1.10**
190	-1.75	-3.91	8.54	3.46	**1.10**
809	1.08	-2.09	7.52	2.06	**1.10**
673	-1.11	0.89	12.70	4.66	**1.10**
1000	-1.16	0.57	12.43	5.36	**1.10**
20	-1.19	-2.19	9.70	4.02	**1.11**
450	1.65	-1.05	8.00	2.51	**1.11**
156	0.60	0.07	10.17	4.40	**1.13**
1063	1.32	-1.43	7.95	2.89	**1.13**
672	0.87	1.70	11.53	4.68	**1.14**
345	1.10	-2.39	7.21	1.87	**1.14**
665	0.88	-1.14	8.68	2.70	**1.16**
435	0.49	-2.14	8.07	2.42	**1.17**
257	1.54	1.79	10.95	4.46	**1.18**

Table C2 (continued).

Monomer Number	Monomer Relative HOMO	Monomer Relative LUMO	Monomer Singlet Gap	Monomer ΔE_{T-S}	Oligomer ΔE_{T-S}
1131	-1.95	-2.64	10.00	4.11	**1.18**
1216	1.50	-2.02	7.18	2.11	**1.18**
538	1.26	1.00	10.43	4.60	**1.18**
440	0.18	-2.36	8.15	3.53	**1.19**
805	0.93	-1.13	8.64	3.10	**1.19**
1072	0.66	1.75	11.79	6.98	**1.19**
694	-0.47	-1.15	10.01	3.69	**1.19**
3	1.79	-1.95	6.95	1.76	**1.19**
117	0.49	-1.09	9.12	3.28	**1.20**
1035	-1.16	0.57	12.43	5.36	**1.21**
1142	-0.02	-0.66	10.06	4.10	**1.21**
521	0.43	1.72	11.98	6.85	**1.23**
1126	-0.57	-2.07	9.19	3.50	**1.26**
416	1.29	1.61	11.02	5.50	**1.28**
206	0.59	-0.91	9.19	3.50	**1.28**
51	0.28	-0.16	10.26	3.79	**1.29**
360	1.07	-0.91	8.72	2.89	**1.29**
151	0.91	-0.41	9.37	3.36	**1.29**
507	0.00	0.00	10.70	4.91	**1.29**
322	-0.25	2.29	13.23	7.44	**1.30**
182	-0.35	0.23	11.27	4.57	**1.30**
1116	0.93	-0.14	9.63	3.69	**1.31**
819	-0.10	-0.49	10.31	4.72	**1.31**
947	1.59	-1.23	7.87	2.65	**1.31**
499	-0.50	-0.36	10.84	4.97	**1.31**
559	0.66	-1.16	8.88	3.03	**1.32**
288	0.97	-0.67	9.06	3.39	**1.32**
932	2.03	-1.35	7.32	2.67	**1.32**
78	0.47	1.11	11.33	4.61	**1.32**
587	1.51	-2.67	6.52	1.31	**1.33**
1094	2.12	1.14	9.72	3.59	**1.33**
766	-0.77	-0.95	10.52	5.43	**1.34**
874	0.97	-1.42	8.30	2.94	**1.34**
698	0.07	-1.53	9.10	3.23	**1.35**
900	-0.45	0.38	11.53	4.46	**1.35**
425	0.65	0.24	10.29	4.24	**1.38**
393	0.97	-1.12	8.61	3.00	**1.38**
542	-0.21	-0.53	10.38	4.74	**1.40**
28	1.07	-2.80	6.83	1.85	**1.41**
467	-0.34	0.44	11.48	4.39	**1.41**
981	1.27	1.53	10.95	4.29	**1.42**
987	0.09	-1.55	9.06	3.48	**1.43**
784	1.52	-1.06	8.11	3.02	**1.43**
555	1.12	1.38	10.95	4.22	**1.43**
1090	-0.86	-3.15	8.41	3.46	**1.44**
444	0.74	-1.78	8.17	2.89	**1.46**
79	0.96	-1.01	8.73	2.89	**1.46**
310	0.15	-1.04	9.51	3.57	**1.47**
350	0.90	-1.06	8.74	3.46	**1.47**
412	0.43	0.15	10.41	3.87	**1.47**
468	-0.35	-2.45	8.60	2.98	**1.48**
1118	1.67	0.01	9.04	3.44	**1.48**
249	0.74	1.55	11.51	5.29	**1.48**
292	0.17	-0.51	10.02	3.90	**1.49**

Table C2 (continued).

Monomer Number	Monomer Relative HOMO	Monomer Relative LUMO	Monomer Singlet Gap	Monomer ΔE_{T-S}	Oligomer ΔE_{T-S}
979	0.50	1.02	11.22	5.43	**1.51**
552	1.67	0.44	9.46	3.58	**1.53**
439	-0.87	0.13	11.70	4.58	**1.55**
724	0.46	1.01	11.25	4.74	**1.59**
473	1.24	-0.34	9.11	3.70	**1.61**
898	-0.34	-2.37	8.67	2.92	**1.62**
687	1.56	-0.69	8.45	2.94	**1.67**
983	1.59	-1.99	7.12	1.93	**1.94**
226	1.51	-0.12	9.06	3.18	**2.02**
659	1.08	-1.59	8.03	2.43	**2.03**

Appendix D QupKake: Integrating Machine Learning and Quantum Chemistry for micro-pKa Predictions

D.1 Code and Data Availability

Full code, data files, and analysis notebooks are available at

`https://github.com/Shualdon/QupKake`

D.2 Supplementary Information

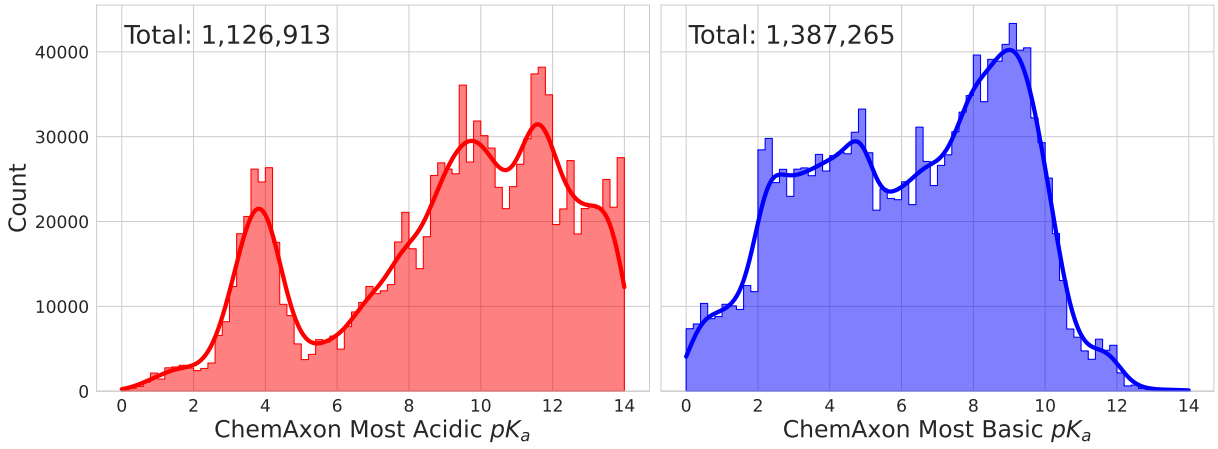


Figure D1: ChemAxon acidic and basic pK_a distribution in the ChEMBL dataset.

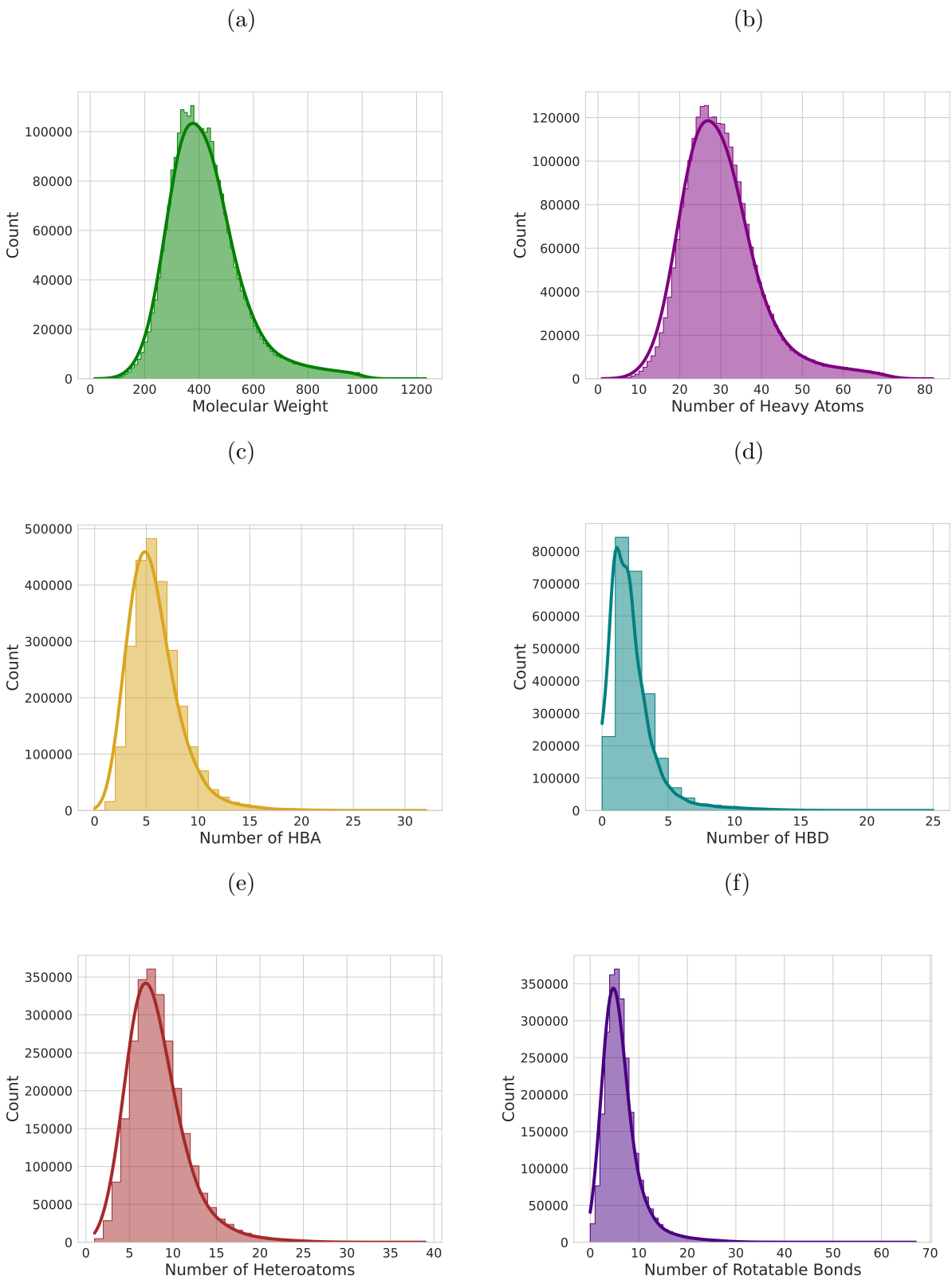


Figure D2: Molecular descriptors for the ChEMBL dataset.

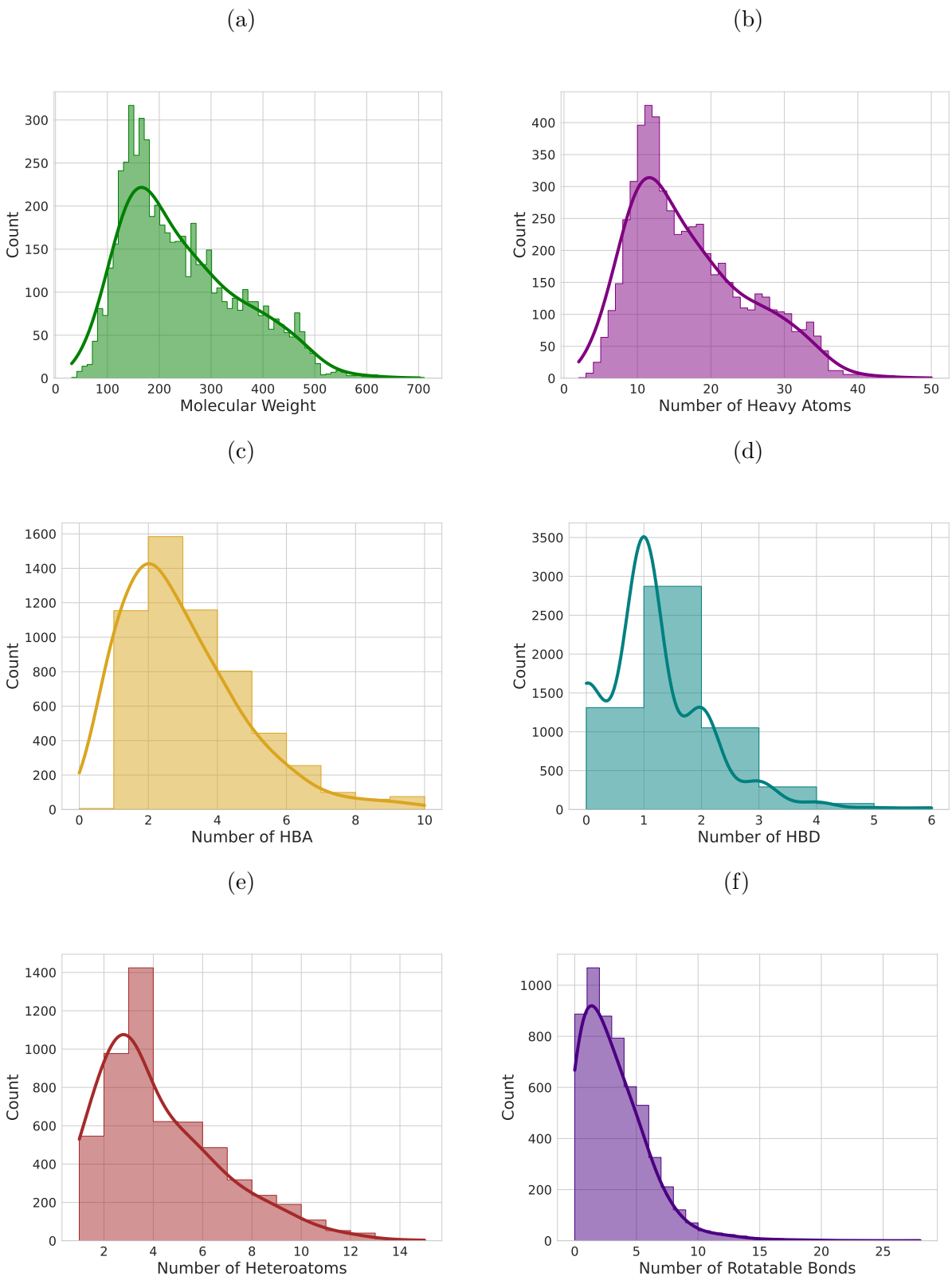


Figure D3: Molecular descriptors for the experimental dataset.

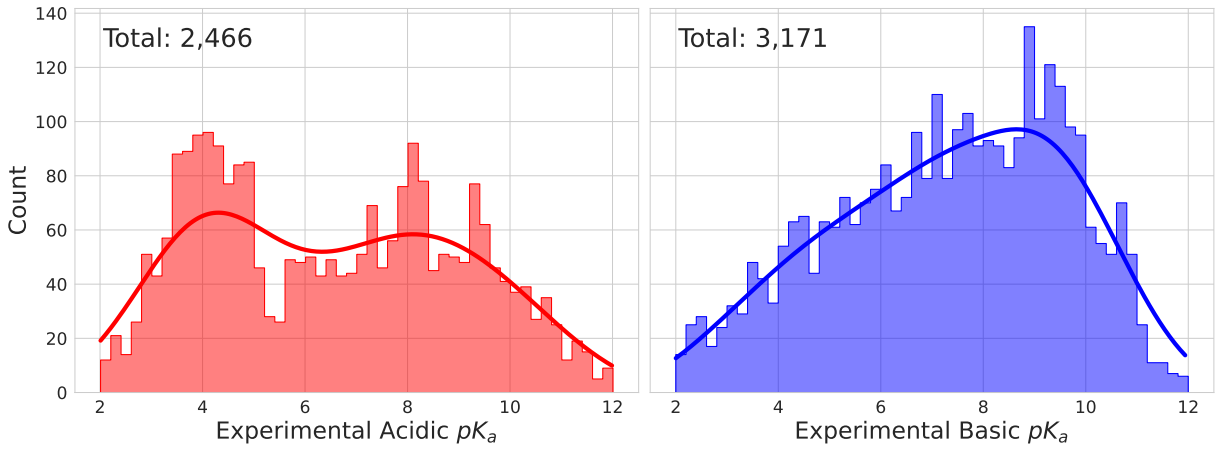


Figure D4: Acidic and basic pK_a distribution in the experimental dataset.

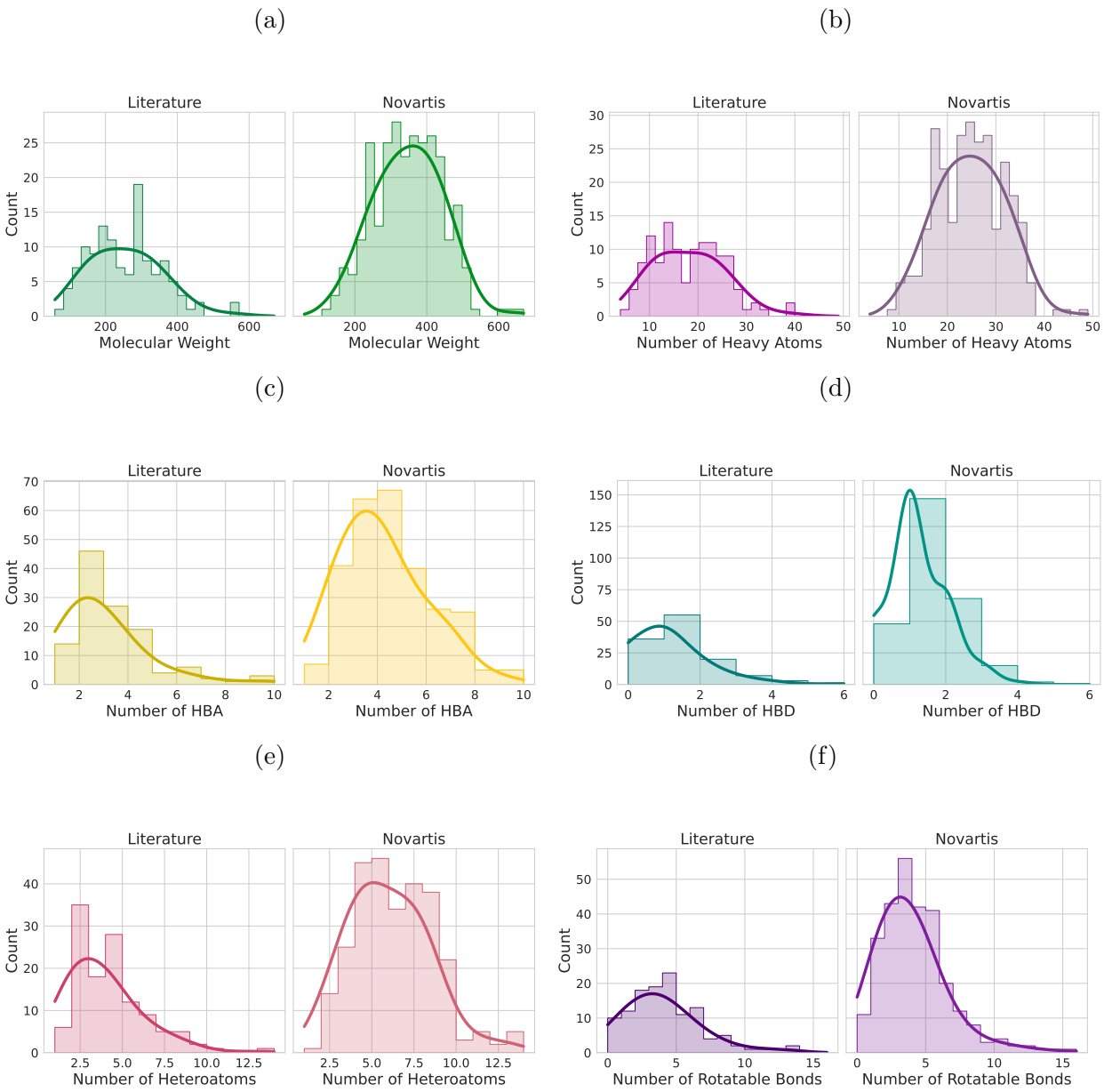


Figure D5: Molecular descriptors for the Literature and Novartis test datasets.

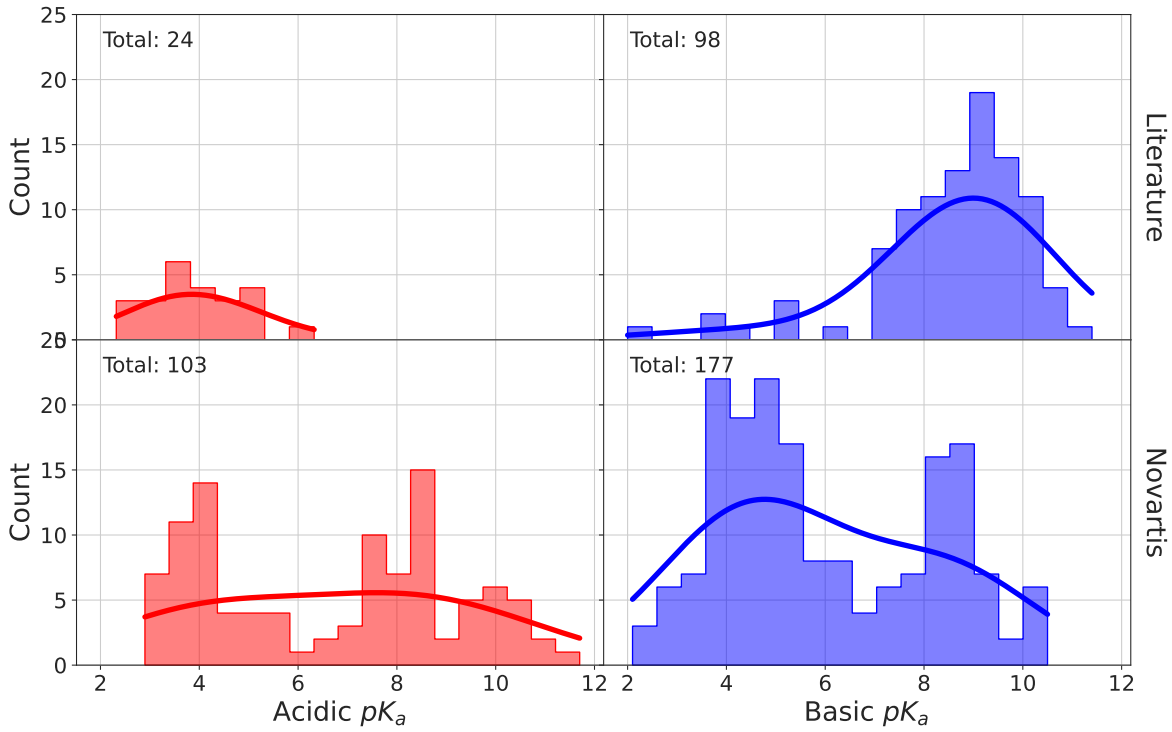
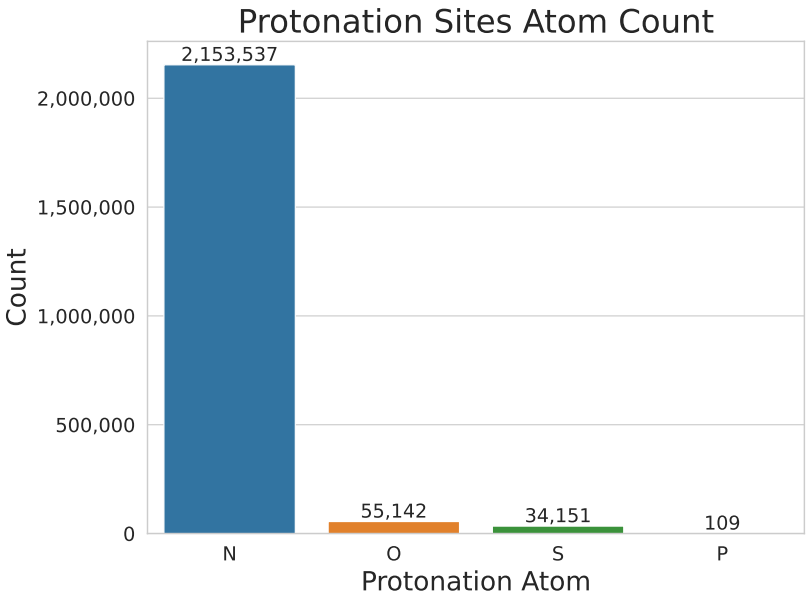


Figure D6: Experimental acidic and basic pK_a distribution in the Litarature (top row) and Novartis (bottom row) test datasets.

(a)



(b)

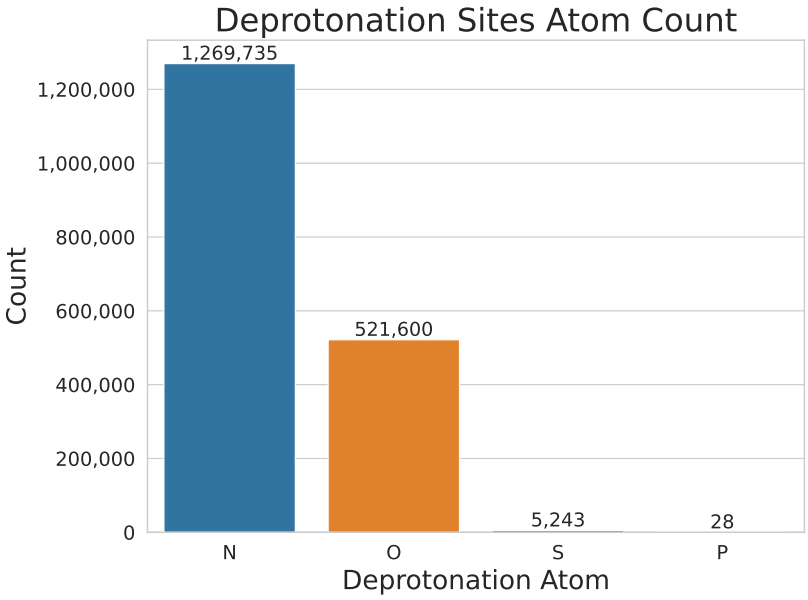


Figure D7: The number of each element that was a protonated or b deprotonated in the CREST datasets.

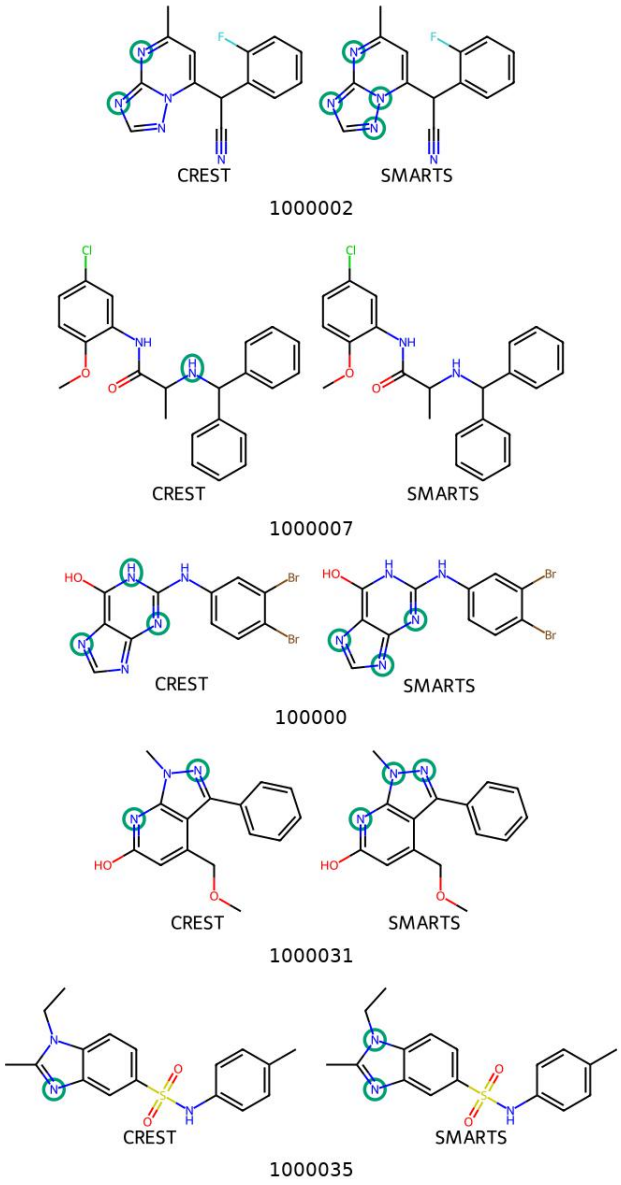


Figure D8: Examples of protonation site discrepancies between CREST and SMARTS patterns. Each row shows the same molecules with the highlighted atoms on the left show the CREST protonation sites while the right shows the SMARTS protonation sites. The Number under each pair of molecules is the ChEMBL ID.

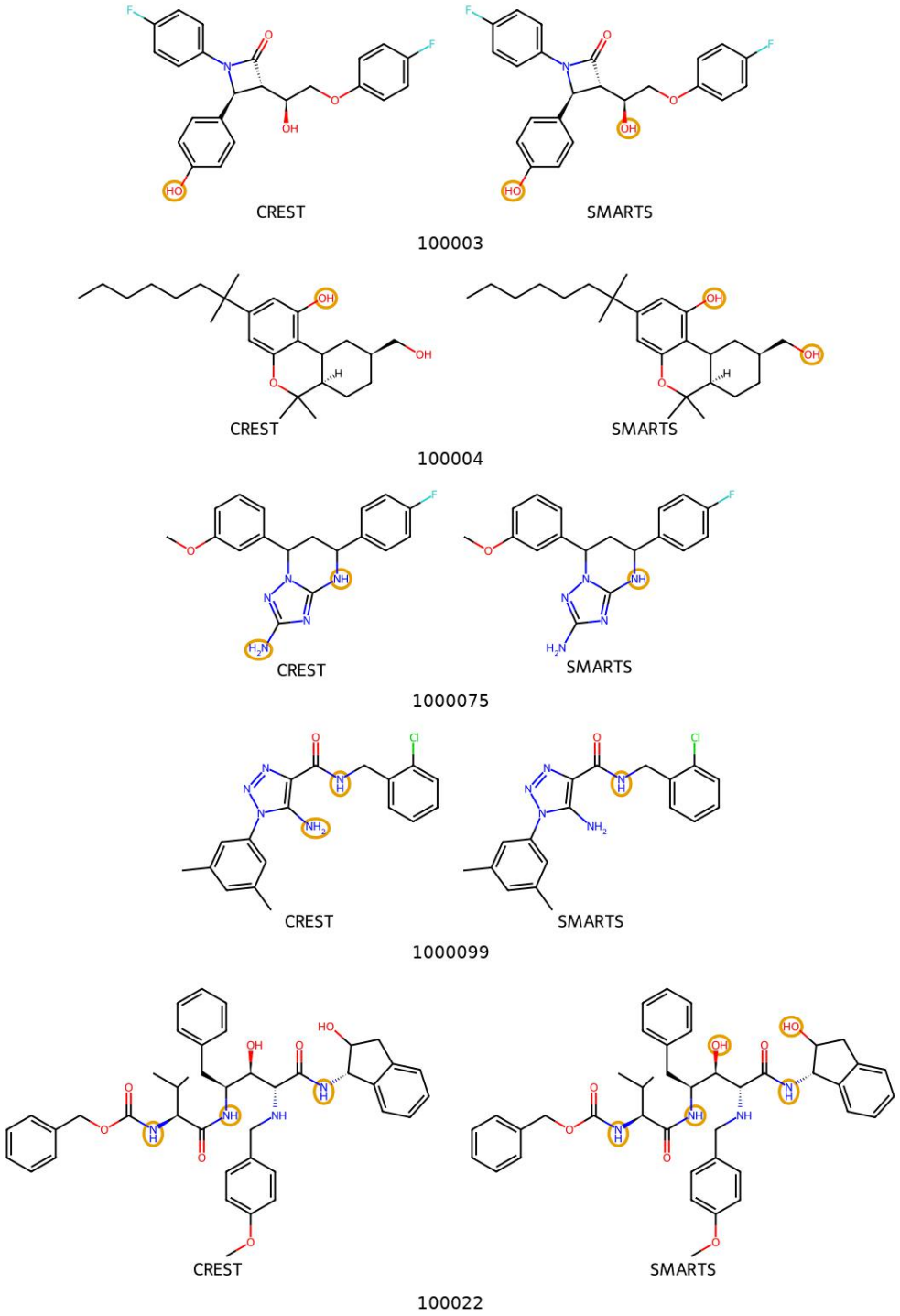


Figure D9: Examples of deprotonation site discrepancies between CREST and SMARTS patterns. Each row shows the same molecules with the highlighted atoms on the left show the CREST deprotonation sites while the right shows the SMARTS deprotonation sites. The Number under each pair of molecules is the ChEMBL ID.

Table D1: The molecular graph features used in the model.

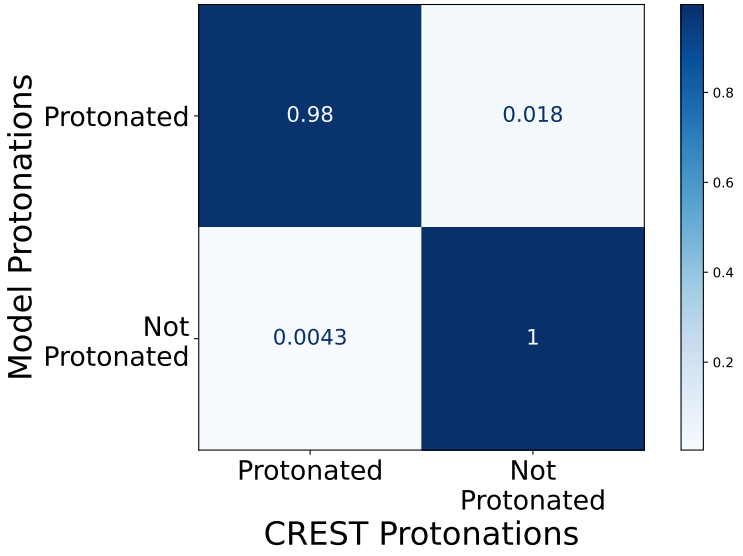
Source	Feature	Value	Length
	Atom Features		
RDKit	Atom Type	One-hot encoding	17
	Heavy Atom Neighbors	One-hot encoding	6
	Formal Charge	One-hot encoding	8
	Hybridization	One-hot encoding	7
	Is In Ring	Binary	1
	Is Aromatic	Binary	1
	Atomic Mass	Float	1
	Van-Der Waals Radius	Float	1
	Covalent Radius	Float	1
	Chirality	One-hot encoding	4
	Number of Hydrogens	One-hot encoding	6
	Is H ydrogen-Bond Donor	Binary	1
	Is Hydrogen-Bond Acceptor	Binary	1
GFN2	Partial Charge	Float	1
	Coordination Number	Float	1
	Polarizability	Float	1
	Fukui Indices	Float	3
	Total		**61**
	Bond Features		
RDKit	Bond Type	One-hot encoding	4
	Is Conjugated	Binary	1
	Is In Ring	Binary	1
	Stereochemistry	One-hot encoding	4
GFN2	Wiberg Bond Order	Float	1
	Total		**11**
	Molecule Features		
RDKit	Radius of Gyration	Float	1
	Spherocity	Float	1
	Aspherocity	Float	1
	Eccentricity	Float	1
	Fraction sp^3 Carbons	Float	1
GFN2	$\Delta E_{ionization}$	Float	1
	Charge	Float	1
	Total		**7**

Table D2: Tuned hyperparameters for the reaction sites models, as found by Optuna. Both protonation and deprotonation models use the same hyperparameters.

Hyperparameter	Possible Values	Selected Value
GNN Architecture	GCNNet, GATNet, TransformetNet	TransformetNet
Num. of Attention Heads †	[1, 2, 3, 4]	2
Hidden Layer Size	64-512	94
Number of GNN Layers	[1, 2, 3, 4]	2

† Num. of Attention Heads parameter is only used for the GATNet and TransformerNet architectures.

(a)



(b)

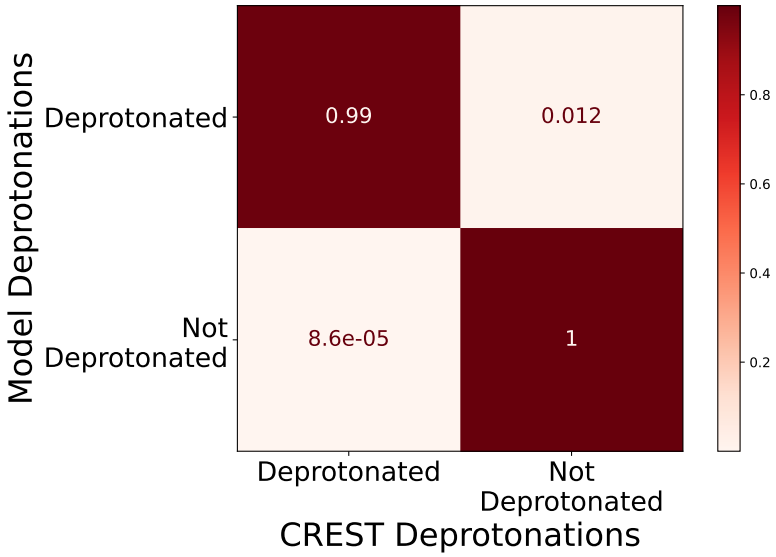
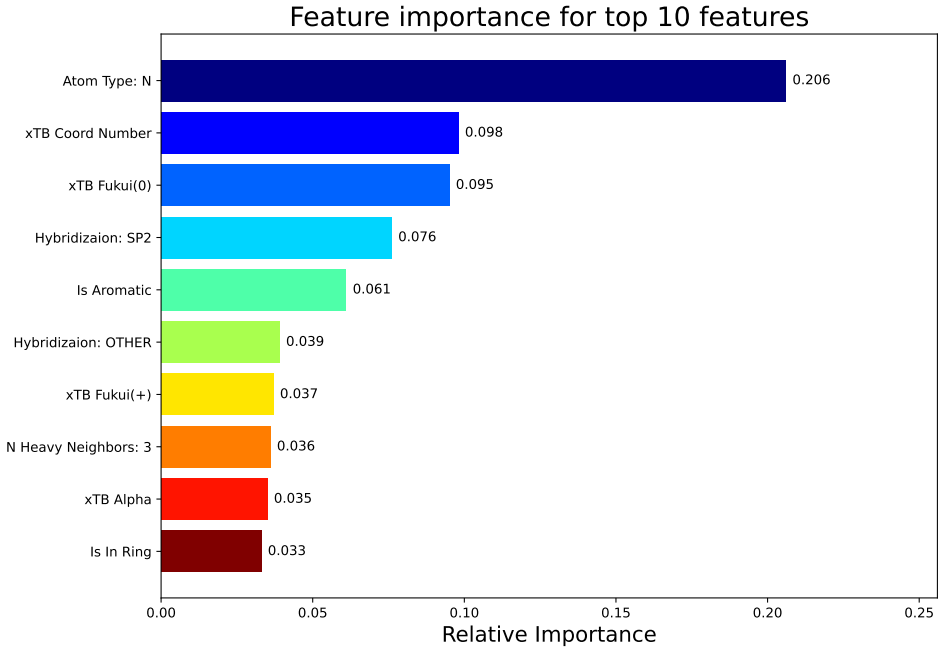


Figure D10: Confusion matrices for the a protonation and b deprotonation reaction sites enumeration models.

(a)



(b)

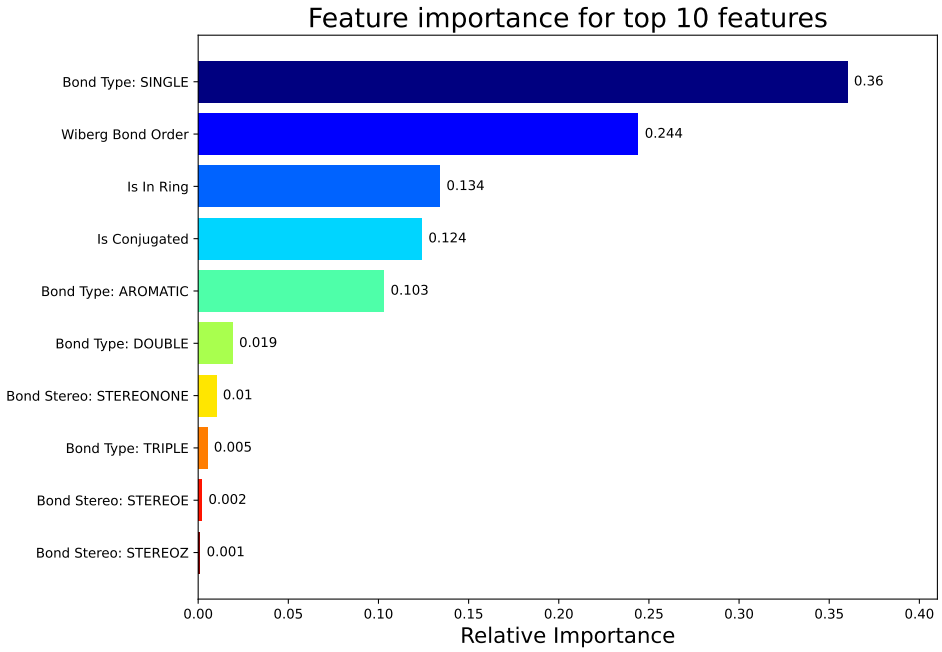
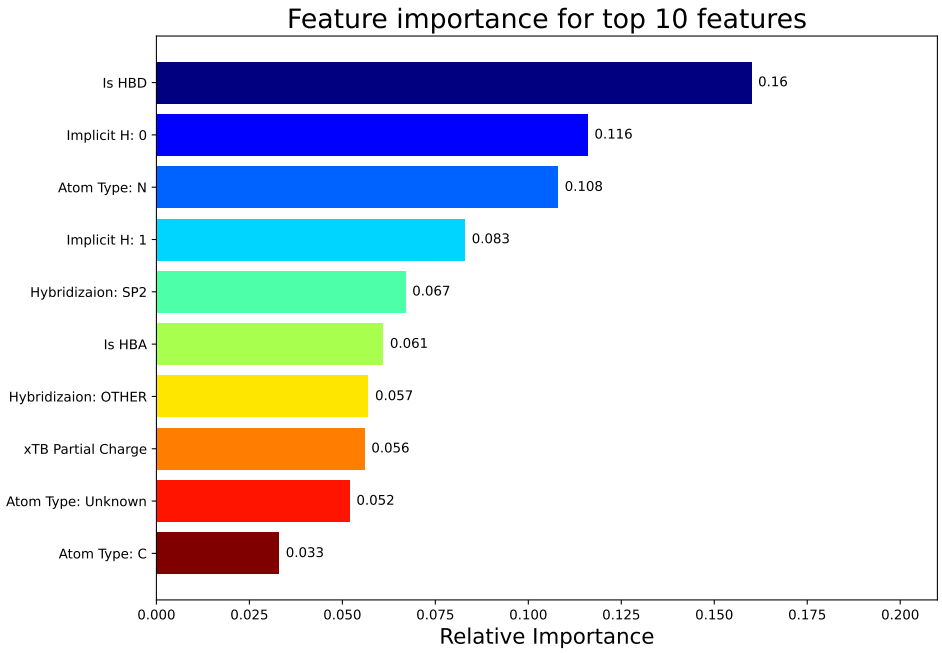


Figure D11: The normalized, absolute relative importance of the a atomic features and b bond features for the protonation sites enumeration model.

(a)



(b)

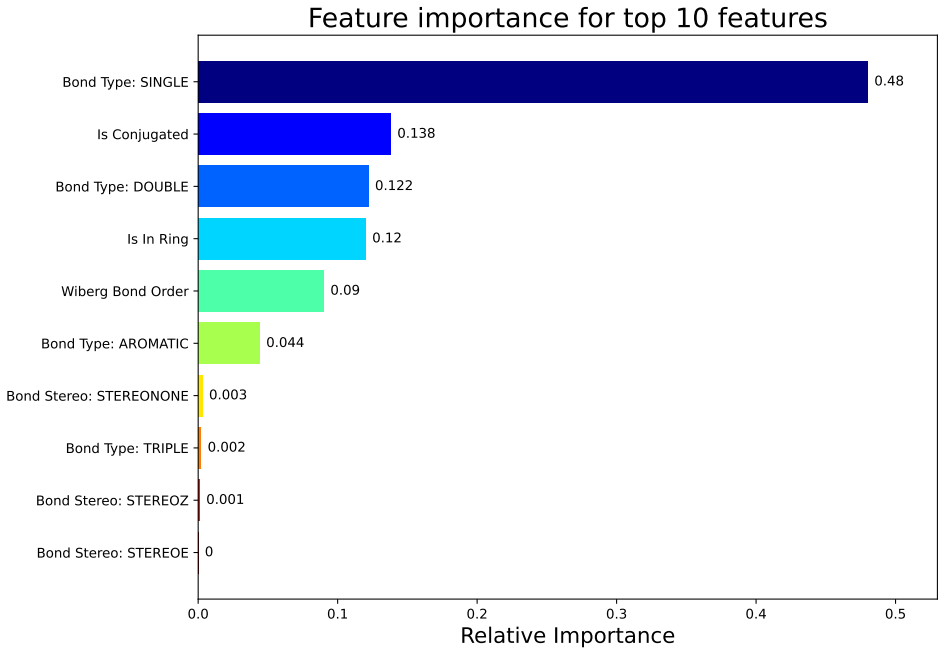


Figure D12: The normalized, absolute relative importance of the a atomic features and b bond features for the deprotonation sites enumeration model.

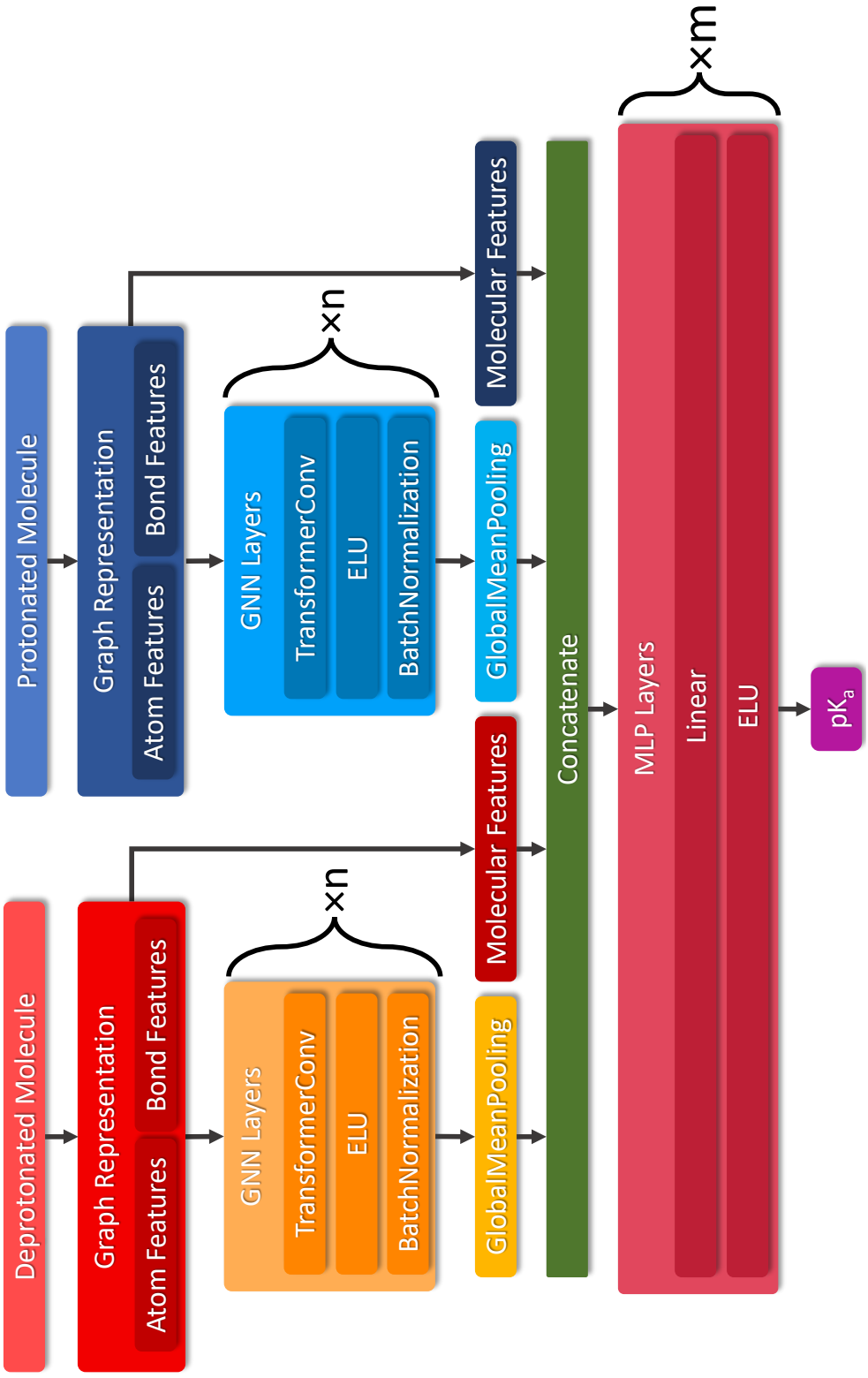


Figure D13: Micro-pK_a prediction model architecture. The parameters n and m correspond to the *Number of GNN Layers* and *Number of MLP Layers*, respectively, in Table D3.

Table D3: Tuned hyperparameters for the micro-pK_a prediction model, as found by Optuna.

Hyperparameter	Possible Values	Selected Value
GNN Architecture	GCNNet, GATNet, TransformetNet	TransformetNet
Number of Attention Heads †	[1, 2, 3, 4]	3
Hidden Layer Size	64-512	51
Number of GNN Layers	[1, 2, 3, 4]	3
Number of MLP Layers	[1, 2, 3, 4]	1

† Number of Attention Heads parameter is only used for the GATNet and TransformerNet architectures.

(a) (b)

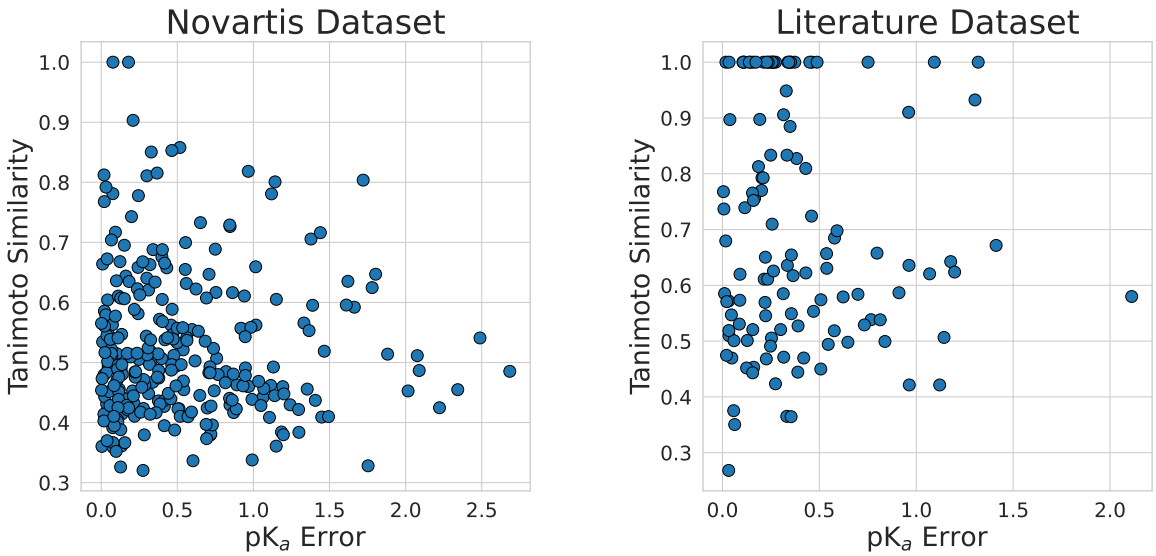


Figure D14: The highest Tanimoto similarity score of the a Novartis test set and the b Experimental test set, compared to the transfer training set, versus the pK_a error, i.e. the absolut difference between the experimantal and predicts pK_a values. As can be seen, there is no obvious correlation, indicating that the existence of a similar molecule in the training set has a negligible effect on the models' pK_a prediction.

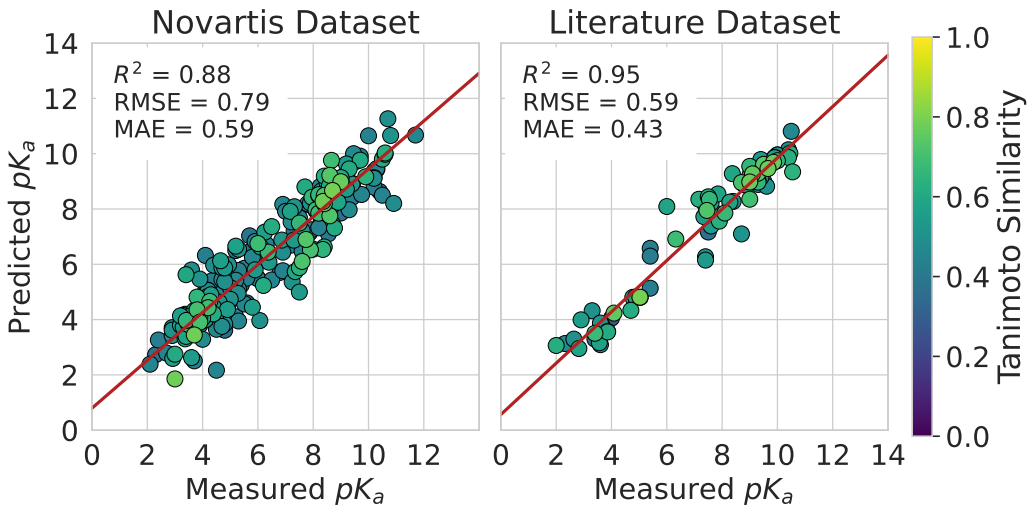


Figure D15: Micro-pK_a predictions versus the measured micro-pK_a values of the Novartis and Literature datasets, filtered to include only molecules with low (< 0.8) Taminoto similarity scores. Data points are colored according to the highest Tanimoto similarity score of the molecule in the test set versus the molecules in the experimental training set. The best-fit linear regression line is shown in red.

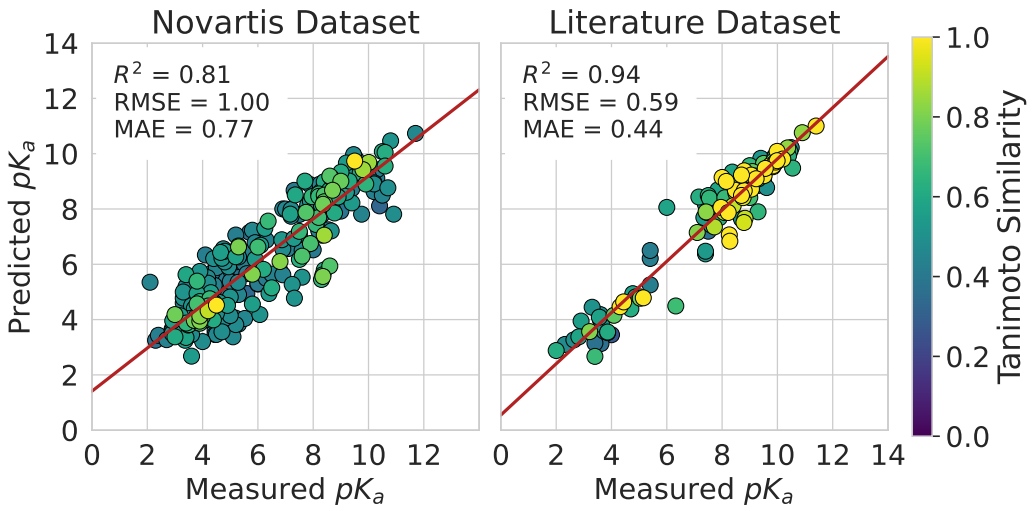


Figure D16: Micro-pK_a predictions versus the measured micro-pK_a values of the Novartis and Literature datasets, using the ChemAxon Marvin predicted reaction centers. Data points are colored according to the highest Tanimoto similarity score of the molecule in the test set versus the molecules in the experimental training set. The best-fit linear regression line is shown in red.

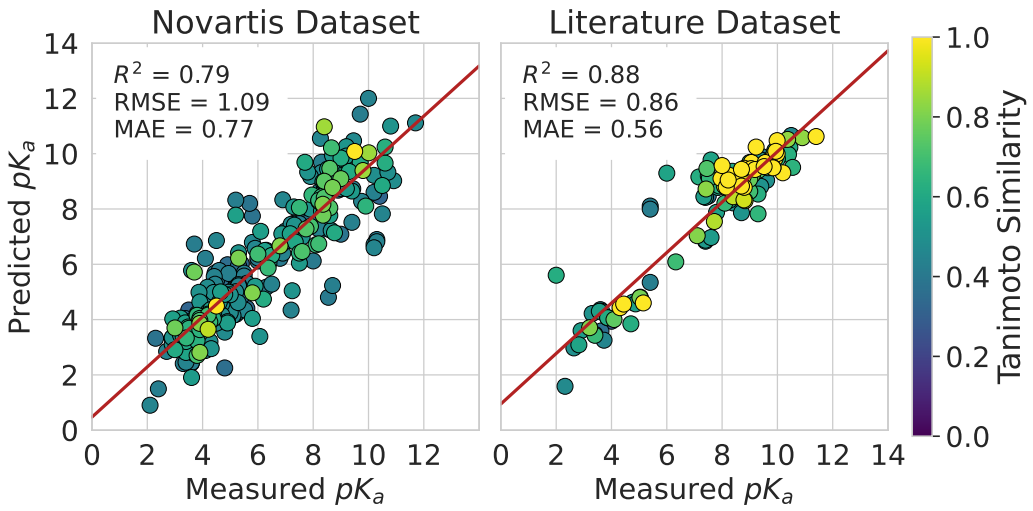


Figure D17: Micro-pK_a predictions versus the measured micro-pK_a values of the Novartis and Literature datasets, using a model trained only on the ChEMBL dataset without transfer learning. Data points are colored according to the highest Tanimoto similarity score of the molecule in the test set versus the molecules in the experimental training set. The best-fit linear regression line is shown in red.

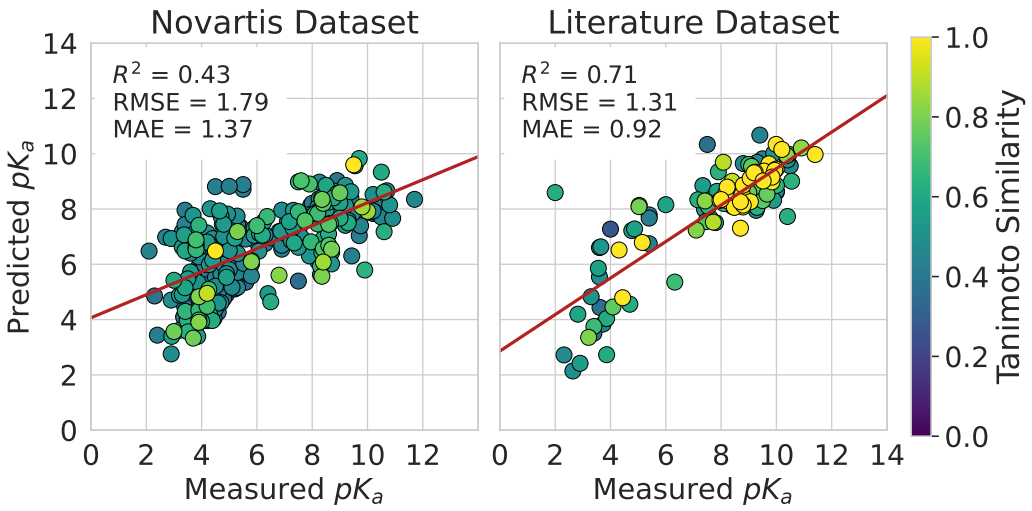


Figure D18: Micro-pK_a predictions versus the measured micro-pK_a values of the Novartis and Literature datasets, using a model trained only on the experimental dataset. Data points are colored according to the highest Tanimoto similarity score of the molecule in the test set versus the molecules in the experimental training set. The best-fit linear regression line is shown in red.

Table D4: Predictions on Nitrogen-containing aromatic heterocycles from the Thapa & Raghavachari **Set-I** dataset[199].

			SMD only		SMD + 1 water		QupKake	
S.N.	SMILES	pK_a^{Exp}	pK_a	ΔpK_a	pK_a	ΔpK_a	pK_a	ΔpK_a
1	c1ccncn1	1.10	-0.596	-1.696	0.439	-0.661	2.161	1.061
2	c1cccnn1	2.10	0.909	-1.191	0.927	-1.173	3.185	1.085
3	c1c(cccn1)Cl	2.84	0.570	-2.270	1.828	-1.012	3.016	0.176
4	c1(ccccn1)OC	3.28	2.060	-1.220	2.627	-0.653	3.147	-0.133
5	c1cc2c(cc1)cnnc2	3.39	2.506	-0.884	2.426	-0.964	3.460	0.070
6	c1cc2c(cc1)cccn2	4.85	3.626	-1.224	4.126	-0.724	4.604	-0.246
7	c1c(cccn1)O	4.86	3.140	-1.720	3.926	-0.934	4.454	-0.406
8	c1ccc(cn1)OC	4.88	3.412	-1.468	4.192	-0.688	4.777	-0.103
9	c1cc2c(cc1)cc1c(c2)nccc1	5.05	3.525	-1.525	4.046	-1.004	4.268	-0.782
10	c1ccccn1	5.17	3.838	-1.332	4.323	-0.847	4.881	-0.289
11	c1cc2c(cc1)cc1c(cccc1)n2	5.60	4.425	-1.175	4.483	-1.117	4.801	-0.799
12	c1c(cccn1)CC	5.70	4.382	-1.318	5.106	-0.594	5.282	-0.418
13	c1c(cccn1)C(C)(C)C	5.82	4.499	-1.321	5.038	-0.782	5.433	-0.387
15	c1cccc(n1)CC	5.97	5.081	-0.889	5.247	-0.723	5.643	-0.327
16	c1cc(ccn1)C(C)(C)C	5.99	4.729	-1.261	5.339	-0.651	5.646	-0.344
17	c12c(cccc1)nc[nH]2	6.00	3.650	-2.350	4.942	-1.058	5.038	-0.962
18	c1cc(ccn1)CC	6.02	4.745	-1.275	5.328	-0.692	5.848	-0.172
19	c1[nH]cc(n1)CO	6.45	4.921	-1.529	5.613	-0.837	5.979	-0.471
20	c1cc(ccn1)OC	6.62	5.326	-1.294	5.558	-1.062	6.243	-0.377
21	c1(cnc[nH]1)C[C@@H](C(=O)[O])NC(=O)C	7.05	6.293	-0.757	7.053	0.003	6.100	-0.950
22	c1(ncc[nH]1)C	7.75	7.148	-0.602	8.010	0.260	7.482	-0.268
	MAE			1.348		0.783		0.468
	MSE			1.41		0.829		0.568
	MaxAbsError			2.350		1.173		1.085

Table D5: Predictions on Aliphatic alcohols from the Thapa & Raghavachari **Set-I** dataset[199].

			SMD only		SMD + 1 water		SMD + 2 waters		SMD + 3 waters		QupKake	
S.N.	SMILES	pK_a^{Exp}	pK_a	ΔpK_a	pK_a	ΔpK_a	pK_a	ΔpK_a	pK_a	ΔpK_a	pK_a	ΔpK_a
1	C(C(O)(C)C)(F)(F)F	11.60	20.692	9.092	18.025	6.425	14.767	3.167	12.591	0.991	2.372	-9.228
2	C(CO)(Cl)(Cl)Cl	12.02	18.854	6.834	15.877	3.857	13.017	0.997	11.466	-0.554	0.592	-11.428
3	C(CO)(F)(F)F	12.43	19.877	7.447	16.966	4.536	14.217	1.787	12.047	-0.383	2.584	-9.846
4	C#CCO	13.55	22.443	8.893	18.858	5.308	16.579	3.029	13.663	0.113	2.663	-10.887
5	C(O)COC	14.80	24.311	9.511	20.673	5.873	17.643	2.843	15.187	0.387	5.995	-8.805
6	C(O)C[C@H](O)C	14.90	25.718	10.818	21.410	6.510	19.038	4.138	15.396	0.496	7.968	-6.932
7	C(O)CCO	15.10	25.600	10.500	21.491	6.391	18.603	3.503	15.528	0.428	5.012	-10.088
8	c1ccc(cc1)CO	15.40	24.674	9.274	20.769	5.369	18.112	2.712	14.968	-0.432	2.425	-12.975
9	C(O)/C=C/C	15.52	25.482	9.962	21.257	5.737	18.422	2.902	15.817	0.297	3.424	-12.096
10	CO	15.54	26.118	10.578	21.838	6.298	18.880	3.340	16.474	0.934	10.409	-5.131
11	C(O)C	15.90	25.961	10.061	21.797	5.897	18.407	2.507	16.095	0.195	9.918	-5.982
12	C(O)CC	16.10	26.072	9.972	21.758	5.658	18.971	2.871	15.961	-0.139	4.261	-11.839
13	C1CCC(CC1)O	16.84	26.543	9.703	22.169	5.329	19.438	2.598	17.280	0.440	3.964	-12.876
14	C(O)(C)(C)C	17.00	26.515	9.515	22.251	5.251	19.432	2.432	17.028	0.028	8.990	-8.010
15	C(O)(C)C	17.10	26.167	9.067	21.903	4.803	19.649	2.549	16.996	-0.104	8.596	-8.504
	MAE			9.415		5.549		2.758		0.395		9.642
	MSE			9.474		5.597		2.846		0.479		0.991
	MaxAbsError			10.818		6.510		4.138		0.991		12.975

Table D6: Predictions on Aliphatic thiols from the Thapa & Raghavachari **Set-I** dataset[199].

			SMD only		SMD + 1 water		SMD + 2 waters		SMD + 3 waters		QupKake	
S.N.	SMILES	pK_a^{Exp}	pK_a	ΔpK_a	pK_a	ΔpK_a	pK_a	ΔpK_a	pK_a	ΔpK_a	pK_a	ΔpK_a
1	C=CCS	7.86	18.450	10.590	14.685	6.825	12.525	4.665	8.698	0.838	4.004	-3.856
2	SCC(=O)OCC	7.95	14.675	6.725	12.120	4.170	9.128	1.178	7.377	-0.573	3.875	-4.075
3	SC[C@@H](CO)S	8.62	17.444	6.874	14.306	3.736	12.747	2.177	7.720	-0.900	4.788	-3.832
4	C(OCC)CS	9.38	17.947	8.567	14.398	5.018	11.887	2.507	8.553	-0.827	3.624	-5.756
5	OCCS	9.72	18.114	8.394	14.788	5.068	12.147	2.427	9.515	-0.205	4.354	-5.366
6	SC(C)(C)CO	9.85	17.777	7.927	13.492	3.642	11.539	1.689	8.971	-0.879	3.793	-6.057
7	C(=C)CS	9.96	17.945	7.985	14.643	4.683	11.992	2.032	9.263	-0.697	3.865	-6.095
8	C(C(=O)[O])CS	10.27	17.468	7.198	13.672	3.402	13.946	3.676	11.332	1.062	3.404	-6.866
9	SC	10.33	19.557	9.227	15.986	5.656	13.348	3.018	10.147	-0.183	7.192	-3.138
10	CCS	10.61	19.589	8.979	16.046	5.436	13.238	2.628	10.545	-0.065	6.536	-4.074
11	C(CC)CS	10.67	19.752	9.082	15.772	5.102	13.533	2.863	10.508	-0.162	4.402	-6.268
12	SC(C)C	10.86	19.516	8.656	16.403	5.543	13.397	2.537	10.695	-0.165	5.795	-5.065
13	SC(C)(C)C	11.05	19.882	8.832	15.999	4.949	13.729	2.679	11.005	-0.045	6.480	-4.570
14	SC(C)(C)CC	11.22	19.994	8.774	16.508	5.288	14.069	2.849	10.770	-0.450	5.037	-6.183
	MAE			8.554		5.033		2.777		0.504		5.086
	MSE			8.601		5.104		2.913		0.612		5.207
	MaxAbsError			10.59		6.825		4.665		1.062		6.866

Table D7: Predictions on primary amines from the Thapa & Raghavachari **Set-I** dataset[199].

			SMD only		SMD + 1 water		SMD + 2 waters		SMD + 3 waters		QupKake	
S.N.	SMILES	pK_a^{Exp}	pK_a	ΔpK_a	pK_a	ΔpK_a	pK_a	ΔpK_a	pK_a	ΔpK_a	pK_a	ΔpK_a
1	NCC#N	5.30	3.245	-2.055	3.510	-1.790	4.896	-0.404	5.883	0.583	5.254	-0.046
2	c1cc(ccc1)CN	9.34	9.029	-0.311	8.914	-0.426	10.114	0.774	11.268	1.928	8.747	-0.593
3	c1cc(ccc1)CCN	9.68	9.138	-0.542	8.728	-0.952	9.656	-0.024	10.694	1.014	9.399	-0.281
4	C(N)(C)C	9.80	10.595	0.795	9.814	0.014	10.431	0.631	11.290	1.490	9.880	0.080
5	NCCCC	10.59	10.334	-0.256	9.463	-1.127	10.269	-0.321	11.268	0.678	10.014	-0.576
6	C(CN)C	10.60	10.481	-0.119	9.725	-0.875	10.466	-0.134	11.547	0.947	9.998	-0.602
7	CN	10.63	10.193	-0.437	9.585	-1.045	10.457	-0.173	10.782	0.152	11.484	0.854
8	C(N)(C)(C)C	10.68	10.833	0.153	9.911	-0.769	10.895	0.215	12.004	1.324	10.336	-0.344
9	C(N)C	10.70	10.258	-0.442	9.668	-1.032	10.622	-0.078	11.098	0.398	9.923	-0.777
10	[C@@H]1(CC[C@@H](CC1)N)C(C)(C)C	11.23	10.879	-0.351	10.165	-1.065	10.643	-0.587	12.900	1.670	10.693	-0.537
	MAE			0.546		0.910		0.334		1.018		0.469
	MSE			0.765		1.010		0.413		1.157		0.536
	MaxAbsError			2.055		1.79		0.774		1.928		0.854

Table D8: Predictions on secondary amines I from the Thapa & Raghavachari **Set-I** dataset[199].

			SMD only		SMD + 1 water		SMD + 2 waters		QupKake	
S.N.	SMILES	pK_a^{Exp}	pK_a	ΔpK_a	pK_a	ΔpK_a	pK_a	ΔpK_a	pK_a	ΔpK_a
1	N(C[C@@H](c1ccc(c(c1)O)O)O)C	8.55	8.446	-0.104	8.925	0.375	10.520	1.970	8.912	0.362
2	C(NC)C	10.54	10.749	0.209	10.658	0.118	12.211	1.671	10.038	-0.502
3	CNC	10.78	10.288	-0.492	10.494	-0.286	11.870	1.090	10.206	-0.574
4	CCCNCCC	11.00	11.106	0.106	10.876	-0.124	12.148	1.148	10.560	-0.440
5	C(NCC)C	11.02	11.161	0.141	10.748	-0.272	12.189	1.169	10.000	-1.020
6	C1CCNCC1	11.22	10.864	-0.356	10.791	-0.429	11.696	0.476	10.489	-0.731
7	C(CNC1CCCCC1)C	11.23	11.286	0.056	11.362	0.132	11.903	0.673	11.897	0.667
8	C1CCCN1	11.27	10.447	-0.823	10.605	-0.665	11.766	0.496	10.943	-0.327
	MAE			0.286		0.300		1.087		0.578
	MSE			0.377		0.348		1.197		0.616
	MaxAbsError			0.823		0.665		1.970		1.020

Table D9: Predictions on secondary amines II from the Thapa & Raghavachari **Set-I** dataset[199].

			SMD only		SMD + 1 water		QupKake	
S.N.	SMILES	pK_a^{Exp}	pK_a	ΔpK_a	pK_a	ΔpK_a	pK_a	ΔpK_a
S.N.	SMILES	pK_a^{Exp}	pK_a	ΔpK_a	pK_a	ΔpK_a	pK_a	ΔpK_a
1	C1CCN(CC1)CC=C	9.69	9.693	0.003	10.160	0.470	9.112	-0.578
2	C1CCN(CC1)CC=C	9.69	9.987	0.297	10.172	0.482	9.071	-0.619
3	CN(C)C	9.80	9.641	-0.159	10.802	1.002	9.126	-0.674
4	C(N(C)C)C	10.16	10.218	0.058	10.797	0.637	9.606	-0.554
5	CCN(CC)CC	10.75	11.490	0.740	11.173	0.423	10.097	-0.653
	MAE			0.251		0.603		0.616
	MSE			0.365		0.639		0.617
	MaxAbsError			0.740		1.002		0.674

Table D10: Predictions on carboxylic acids from the Thapa & Raghavachari **Set-I** dataset[199].

			SMD only		SMD + 1 water		SMD + 2 waters		SMD + 3 waters		QupKake	
S.N.	SMILES	pK_a^{Exp}	pK_a	ΔpK_a	pK_a	ΔpK_a	pK_a	ΔpK_a	pK_a	ΔpK_a	pK_a	ΔpK_a
1	C(=O)(O)C(F)(F)F	-0.26	-1.526	-1.266	-0.088	0.172	-0.267	-0.007	-0.668	-0.408	2.101	2.361
2	C(=O)(O)C(Cl)(Cl)Cl	0.65	-1.123	-1.772	-0.014	-0.664	-0.015	-0.665	-0.817	-1.467	1.445	0.795
3	C(=O)(O)C(F)F	1.24	0.914	-0.326	1.791	0.551	1.921	0.681	0.893	-0.347	2.129	0.889
4	C(=O)(O)C(Cl)Cl	1.30	1.126	-0.174	1.924	0.624	1.539	0.239	0.705	-0.595	0.655	-0.645
5	C(#N)CC(=O)O	2.44	2.893	0.453	3.227	0.787	3.138	0.698	2.202	-0.238	2.239	-0.201
6	C(=O)(O)CF	2.66	3.533	0.873	3.726	1.066	3.266	0.606	2.615	-0.045	1.933	-0.727
7	C(=O)(O)[C@@H](Cl)C	2.80	4.636	1.836	4.276	1.476	4.235	1.435	3.114	0.314	1.666	-1.134
8	C(=O)(O)CCl	2.81	4.434	1.624	4.625	1.815	4.395	1.585	2.968	0.158	2.036	-0.774
9	C#CCC(=O)O	2.86	4.748	1.888	4.786	1.926	4.325	1.465	3.654	0.794	2.300	-0.560
10	C(=O)(O)CBr	2.86	4.094	1.234	4.367	1.507	3.871	1.011	2.74	-0.12	2.306	-0.554
11	C(=O)(O)CC(F)(F)F	3.07	4.650	1.580	4.698	1.628	4.442	1.372	3.081	0.011	2.922	-0.148
12	C(=O)(O)CC(=O)C	3.53	5.244	1.714	5.022	1.492	4.702	1.172	3.391	-0.139	2.113	-1.417
13	COCC(=O)O	3.54	4.642	1.102	4.747	1.207	4.315	0.775	3.563	0.023	2.198	-1.342
14	C(=O)O	3.75	5.333	1.583	4.774	1.024	4.658	0.908	2.707	-1.043	0.144	-3.606
15	C(=O)(O)CO	3.83	4.835	1.005	5.149	1.319	4.395	0.565	3.567	-0.263	2.206	-1.624
16	C(=O)(O)[C@H](O)C	3.87	3.607	-0.263	3.540	-0.330	3.491	-0.379	NaN	NaN	2.143	-1.727
17	C(CC(=O)O)Cl	4.10	6.168	2.068	5.697	1.597	4.843	0.743	4.361	0.261	2.083	-2.017
18	C(=O)(O)C=C	4.26	6.356	2.096	6.120	1.860	5.490	1.230	4.879	0.619	1.168	-3.092
19	c1ccccc1CC(=O)O	4.31	6.279	1.969	5.879	1.569	5.505	1.195	4.153	-0.157	1.802	-2.508
20	C=CCC(=O)O	4.35	6.376	2.026	6.035	1.685	5.337	0.987	4.777	0.427	3.397	-0.953
21	ClCCCC(=O)O	4.52	6.969	2.449	6.490	1.970	5.764	1.244	4.654	0.134	1.928	-2.592
22	C(=O)(O)C	4.76	7.718	2.958	7.173	2.413	6.896	2.136	5.378	0.618	2.254	-2.506
23	C(=O)(O)CCC	4.82	8.096	3.276	7.204	2.384	6.906	2.086	5.557	0.737	1.974	-2.846
24	C(=O)(O)CC	4.87	8.107	3.237	7.552	2.682	7.200	2.330	5.594	0.724	2.017	-2.853
25	C1CCC(CC1)C(=O)O	4.90	7.953	3.053	7.356	2.456	6.812	1.912	5.542	0.642	2.508	-2.392
26	C(=O)(O)C(C)(C)C	5.05	8.059	3.009	7.616	2.566	7.103	2.053	5.917	0.867	2.188	-2.862
	MAE			1.724		1.491		1.134		0.446		1.659
	MSE			1.943		1.642		1.285		0.569		1.931
	MaxAbsError			3.276		2.682		2.330		1.467		3.606

Table D11: Predictions on thiophenols from the Thapa & Raghavachari **Set-I** dataset[199].

			SMD only		SMD + 1 water		SMD + 2 waters		SMD + 3 waters		QupKake	
S.N.	SMILES	pK_a^{Exp}	pK_a	ΔpK_a	pK_a	ΔpK_a	pK_a	ΔpK_a	pK_a	ΔpK_a	pK_a	ΔpK_a
1	Sc1ccc(cc1)N(=O)=O	4.72	8.622	3.902	6.254	1.534	3.743	-0.977	2.276	-2.444	2.340	-2.380
2	Sc1cc(ccc1)N(=O)=O	5.24	9.979	4.739	7.252	2.012	5.204	-0.036	2.977	-2.263	2.477	-2.763
3	Sc1c[nH]c(=O)[nH]c1=O	5.30	10.010	4.710	7.428	2.128	5.220	-0.080	2.689	-2.611	4.437	-0.863
4	Sc1ccc(cc1)C(=O)C	5.33	10.345	5.015	7.826	2.496	5.302	-0.028	3.331	-1.999	2.315	-3.015
5	Sc1cc(ccc1)Cl	5.78	11.171	5.391	8.372	2.592	5.707	-0.073	3.632	-2.148	2.582	-3.198
6	Sc1ccc(cc1)Br	6.02	11.629	5.609	8.835	2.815	6.797	0.777	4.350	-1.670	2.697	-3.323
7	Sc1ccc(cc1)Cl	6.14	11.676	5.536	9.080	2.940	6.797	0.657	4.169	-1.971	2.645	-3.495
8	c1(cc(ccc1)OC)S	6.39	12.533	6.143	9.568	3.178	7.271	0.881	4.746	-1.644	2.839	-3.551
9	Sc1ccccc1	6.61	12.643	6.033	9.933	3.323	7.225	0.615	5.170	-1.440	2.931	-3.679
10	Sc1c(cccc1)C	6.64	13.467	6.827	10.465	3.825	8.328	1.688	5.832	-0.808	2.941	-3.699
11	Sc1cc(ccc1)C	6.66	12.976	6.316	10.115	3.455	7.603	0.943	5.278	-1.382	2.860	-3.800
12	Sc1cc(ccc1)OC	6.78	12.500	5.720	9.661	2.881	7.142	0.362	4.992	-1.788	2.837	-3.943
13	Sc1ccc(cc1)C	6.82	13.226	6.406	10.454	3.634	7.798	0.978	5.557	-1.263	3.142	-3.678
	MAE			5.565		2.832		0.623		1.802		3.184
	MSE			5.620		2.904		0.784		1.867		3.281
	MaxAbsError			6.827		3.825		1.688		2.611		3.943

Table D12: Predictions on phenols from the Thapa & Raghavachari **Set-I** dataset[199].

			SMD only		SMD + 1 water		SMD + 2 waters		SMD + 3 waters		QupKake	
S.N.	SMILES	pK_a^{Exp}	pK_a	ΔpK_a	pK_a	ΔpK_a	pK_a	ΔpK_a	pK_a	ΔpK_a	pK_a	ΔpK_a
1	c1cc(c(cc1)O)C=O	6.79	14.197	7.407	11.580	4.790	10.508	3.718	6.894	0.104	2.029	-4.761
2	c1(ccc(cc1)O)N(=O)=O	7.14	9.699	2.559	9.068	1.928	7.625	0.485	5.877	-1.263	2.321	-4.819
3	c1ccc(c(c1)N(=O)=O)O	7.23	8.498	1.268	9.411	2.181	7.253	0.023	6.013	-1.217	2.492	-4.738
4	c1c(ccc(c1)O)C=O	7.66	11.415	3.755	10.388	2.728	8.728	1.068	7.765	0.105	2.734	-4.926
5	c1c(ccc(c1)O)C#N	7.95	11.838	3.888	10.968	3.018	9.043	1.093	7.730	-0.220	1.598	-6.352
6	c1c(cc(cc1)O)C=O	8.00	14.010	6.010	12.135	4.135	10.374	2.374	8.302	0.302	1.781	-6.219
7	c1ccc(cc1N(=O)=O)O	8.35	12.447	4.097	11.516	3.166	8.966	0.616	7.285	-1.065	2.105	-6.245
8	c1cc(ccc1O)C(=O)OC/C=C/C	8.41	12.275	3.865	11.238	2.828	9.425	1.015	7.604	-0.806	1.989	-6.421
9	c1cc(ccc1O)C(=O)OC	8.47	12.299	3.829	11.295	2.825	9.270	0.800	7.870	-0.600	1.938	-6.532
10	c1cc(ccc1O)C(=O)OCCCC	8.47	12.378	3.908	11.255	2.785	9.372	0.902	8.192	-0.278	1.280	-7.190
11	Clc1ccccc1O	8.48	12.777	4.297	11.166	2.686	9.496	1.016	7.196	-1.284	1.760	-6.720
12	c1cc(ccc1O)C(=O)OCC	8.50	12.397	3.897	11.217	2.717	8.970	0.470	8.074	-0.426	1.739	-6.761
13	c1cc(cc(c1)O)C#N	8.61	13.095	4.485	11.859	3.249	9.884	1.274	8.245	-0.365	1.394	-7.216
14	c1ccc(c(c1)F)O	8.81	13.078	4.268	11.620	2.810	9.522	0.712	7.400	-1.410	1.510	-7.300
15	c1ccc(cc1Cl)O	9.02	13.634	4.614	11.814	2.794	9.753	0.733	8.063	-0.957	1.772	-7.248
16	c1ccc(cc1F)O	9.28	13.914	4.634	12.230	2.950	10.168	0.888	8.861	-0.419	1.854	-7.426
17	c1(ccc(cc1)O)Cl	9.38	14.332	4.952	13.005	3.625	10.720	1.340	9.147	-0.233	1.719	-7.661
18	c1(ccc(cc1)O)C(=O)[O]	9.39	14.885	5.495	13.626	4.236	11.218	1.828	9.553	0.163	2.642	-6.748
19	c1ccc(cc1O)O	9.44	14.985	5.545	13.233	3.793	10.986	1.546	9.373	-0.067	2.117	-7.323
20	c1ccc(c(c1)O)O	9.48	12.734	3.254	12.835	3.355	11.386	1.906	9.725	0.245	2.155	-7.325
21	c1c(ccc(c1)c1ccccc1)O	9.51	15.087	5.577	13.175	3.665	10.921	1.411	9.372	-0.138	0.033	-9.477
22	c1cc(cc(c1)c1ccccc1)O	9.59	15.351	5.761	13.279	3.689	10.889	1.299	9.509	-0.081	0.384	-9.206
23	c1ccc(cc1OC)O	9.65	15.103	5.453	13.169	3.519	10.977	1.327	9.242	-0.408	1.720	-7.930
24	c1cc(ccc1O)CO	9.82	15.006	5.186	13.407	3.587	11.382	1.562	9.496	-0.324	1.848	-7.972
25	c1cc(cc(c1)O)CO	9.83	15.149	5.319	13.400	3.570	10.939	1.109	9.277	-0.553	1.476	-8.354
26	c1cc(cc(c1)O)CC	9.90	15.668	5.768	13.909	4.009	11.730	1.830	9.848	-0.052	0.970	-8.930
27	c1cc(c(cc1)O)CO	9.92	13.256	3.336	13.192	3.272	10.876	0.956	9.231	-0.689	1.705	-8.215
28	c1ccc(c(c1)c1ccccc1)O	9.93	15.207	5.277	13.503	3.573	11.573	1.643	9.869	-0.061	0.485	-9.445
29	c1ccc(c(c1)OC)O	9.93	15.629	5.699	13.288	3.358	11.179	1.249	8.878	-1.052	2.149	-7.781
30	c1ccc(cc1C(=O)[O])O	9.94	15.985	6.045	14.187	4.247	11.909	1.969	9.747	-0.193	1.593	-8.347
31	c1(ccc(cc1)O)F	9.95	15.276	5.326	13.498	3.548	10.971	1.021	9.556	-0.394	1.213	-8.737
32	c1(ccc(cc1)O)O	9.96	16.398	6.438	14.705	4.745	12.747	2.787	10.610	0.650	2.391	-7.569
33	c1ccc(cc1)O	9.98	15.444	5.464	13.553	3.573	11.260	1.280	9.601	-0.379	1.210	-8.770
34	c1cc(ccc1O)CC	10.00	15.965	5.965	14.214	4.214	12.030	2.030	10.414	0.414	0.632	-9.368
35	c1ccc(cc1C)O	10.08	15.693	5.613	13.906	3.826	11.476	1.396	10.102	0.022	1.650	-8.430
36	c1(ccc(cc1)O)C	10.19	16.015	5.825	14.441	4.251	12.162	1.972	10.292	0.102	1.347	-8.843
37	c1(ccc(cc1)O)OC	10.20	16.298	6.098	14.752	4.552	12.361	2.161	10.453	0.253	1.804	-8.396
38	c1cc(c(cc1)O)CC	10.20	15.824	5.624	14.024	3.824	11.874	1.674	9.924	-0.276	1.775	-8.425
39	c1ccc(c(c1)C)O	10.28	15.613	5.333	13.875	3.595	11.566	1.286	9.809	-0.471	1.760	-8.520
	MAE			4.901		3.467		1.379		0.463		7.504
	MSE			5.036		3.529		1.534		0.602		7.614
	MaxAbsError			7.407		4.790		3.718		1.410		9.477

Table D13: Predictions on anilines from the Thapa & Raghavachari **Set-I** dataset[199].

			SMD only		SMD + 1 water		SMD + 2 waters		SMD + 3 waters		QupKake	
S.N.	SMILES	pK_a^{Exp}	pK_a	ΔpK_a	pK_a	ΔpK_a	pK_a	ΔpK_a	pK_a	ΔpK_a	pK_a	ΔpK_a
1	c1ccc(c(c1)N(=O)=O)N	0.28	-4.551	-4.831	-2.201	-2.481	-1.370	-1.650	-1.373	-1.653	1.242	0.962
2	c1(ccc(cc1)N)N(=O)=O	0.98	-2.883	-3.863	-1.253	-2.233	0.137	-0.843	1.220	0.240	2.499	1.519
3	c1ccc(c(c1)C(=O)O)N	2.04	-0.664	-2.704	0.480	-1.560	1.671	-0.369	1.671	-0.369	2.620	0.580
4	c1(ccccc1N)C(=O)OCC	2.10	-0.023	-2.123	0.660	-1.440	1.975	-0.125	1.975	-0.125	2.945	0.845
5	c1(ccccc1N)C(=O)OC	2.16	-0.150	-2.310	0.436	-1.724	1.770	-0.390	1.793	-0.367	3.015	0.855
6	c1(ccc(cc1)N)C(=O)OC	2.30	-0.602	-2.902	0.561	-1.739	1.582	-0.718	2.388	0.088	2.892	0.592
7	c1(ccc(cc1)N)C(=O)O	2.32	-1.164	-3.484	0.006	-2.314	1.540	-0.780	2.722	0.402	2.666	0.346
8	c1(ccc(cc1)N)C(=O)OCC	2.38	-0.747	-3.127	0.407	-1.973	1.394	-0.986	2.545	0.165	2.896	0.516
9	c1ccc(cc1N(=O)=O)N	2.45	-1.118	-3.568	-0.115	-2.565	0.851	-1.599	2.013	-0.437	2.383	-0.067
10	Clc1ccccc1N	2.62	-1.305	-3.925	-0.450	-3.070	1.090	-1.530	3.215	0.595	2.966	0.346
11	c1ccc(c(c1)F)N	2.96	-0.541	-3.501	0.011	-2.949	1.359	-1.601	2.957	-0.003	3.084	0.124
12	c1ccc(cc1C(=O)O)N	3.05	0.209	-2.841	1.032	-2.018	2.107	-0.943	3.294	0.244	3.621	0.571
13	c1ccc(cc1Cl)N	3.32	0.232	-3.088	1.210	-2.110	2.635	-0.685	3.640	0.320	3.726	0.406
14	c1ccc(cc1F)N	3.38	0.620	-2.760	1.485	-1.895	2.313	-1.067	3.257	-0.123	3.396	0.016
15	c1(cc(ccc1)N)C(=O)OC	3.56	0.566	-2.994	1.228	-2.332	2.132	-1.428	3.211	-0.349	3.914	0.354
16	c1ccc(c(c1)c1ccccc1)N	3.78	1.126	-2.654	1.557	-2.223	3.198	-0.582	5.029	1.249	3.746	-0.034
17	c1(ccc(cc1)N)Cl	3.81	0.796	-3.014	1.713	-2.097	3.144	-0.666	3.951	0.141	3.867	0.057
18	c1(cccc(c1)N)SC	4.05	1.487	-2.563	1.719	-2.331	3.328	-0.722	4.251	0.201	4.123	0.073
19	c1ccc(cc1O)N	4.17	1.592	-2.578	2.172	-1.998	3.502	-0.668	4.042	-0.128	3.697	-0.473
20	c1(cc(ccc1)N)OCC	4.17	1.722	-2.448	2.339	-1.831	3.187	-0.983	3.858	-0.312	4.056	-0.114
21	c1ccc(cc1OC)N	4.20	1.473	-2.727	2.344	-1.856	3.560	-0.640	4.271	0.071	4.062	-0.138
23	c1ccc(c(c1)C)N	4.38	2.168	-2.212	2.455	-1.925	4.150	-0.230	4.522	0.142	4.533	0.153
24	c1(ccc(cc1)N)SC	4.40	2.101	-2.299	2.820	-1.580	3.730	-0.670	5.081	0.681	4.434	0.034
25	c1(c(cccc1)N)OCC	4.47	1.762	-2.708	1.926	-2.544	2.829	-1.641	5.108	0.638	3.873	-0.597
26	c1ccc(c(c1)OC)N	4.49	1.747	-2.743	1.688	-2.802	2.746	-1.744	4.418	-0.072	3.635	-0.855
27	c1(ccc(cc1)N)F	4.52	1.542	-2.978	2.416	-2.104	3.413	-1.107	4.559	0.039	3.593	-0.927
29	c1ccc(cc1C)N	4.67	2.080	-2.590	2.736	-1.934	3.883	-0.787	4.700	0.030	4.643	-0.027
30	c1ccc(c(c1)O)N	4.72	0.075	-4.645	1.924	-2.796	3.276	-1.444	4.223	-0.497	3.600	-1.120
31	c1(ccc(cc1)N)C	5.07	2.459	-2.611	2.920	-2.150	4.198	-0.872	5.317	0.247	4.603	-0.467
32	c1(ccc(cc1)N)OCC	5.25	2.745	-2.505	3.138	-2.112	4.266	-0.984	5.557	0.307	5.304	0.054
33	c1(ccc(cc1)N)OC	5.29	2.824	-2.466	3.344	-1.946	4.430	-0.860	5.655	0.365	5.138	-0.152
34	c1(ccc(cc1)N)O	5.50	2.759	-2.741	3.311	-2.189	4.002	-1.498	5.176	-0.324	4.872	-0.628
	MAE			2.953		2.151		0.963		0.341		0.438
	MSE			3.021		2.186		1.058		0.481		0.577
	MaxAbsError			4.831		3.070		1.744		1.653		1.519

Table D14: Predictions on benzoic acids from the Thapa & Raghavachari **Set-I** dataset[199].

			SMD only		SMD + 1 water		SMD + 2 waters		SMD + 3 waters		QupKake	
S.N.	SMILES	pK_a^{Exp}	pK_a	ΔpK_a	pK_a	ΔpK_a	pK_a	ΔpK_a	pK_a	ΔpK_a	pK_a	ΔpK_a
1	c1ccc(c(c1)N(=O)=O)C(=O)O	2.17	2.475	0.305	3.435	1.265	3.242	1.072	2.374	0.204	2.104	-0.066
2	Clc1ccccc1C(=O)O	2.94	3.776	0.836	3.723	0.783	3.678	0.738	2.413	-0.527	2.199	-0.741
3	c1ccc(c(c1)C(=O)O)C(=O)O	2.95	4.576	1.626	4.360	1.410	4.411	1.461	3.399	0.449	3.842	0.892
4	c1ccc(c(c1)O)C(=O)O	2.98	5.421	2.441	2.454	-0.526	2.135	-0.845	3.954	0.974	2.799	-0.181
5	c1ccc(c(c1)F)C(=O)O	3.27	4.878	1.608	2.519	-0.751	4.528	1.258	3.488	0.218	1.932	-1.338
6	c1ccc(cc1N(=O)=O)C(=O)O	3.45	4.599	1.149	4.763	1.313	4.290	0.840	3.483	0.033	2.461	-0.989
7	c1(c(cccc1)C(=O)O)C(C)(C)C	3.46	5.732	2.272	5.289	1.829	4.743	1.283	3.468	0.008	2.490	-0.970
8	c1(ccc(cc1)C(=O)O)C(=O)O	3.51	5.364	1.854	5.034	1.524	5.390	1.880	3.779	0.269	3.704	0.194
9	c1c(cc(cc1)C(=O)O)C(=O)O	3.54	5.374	1.834	5.441	1.901	5.537	1.997	3.692	0.152	3.623	0.083
10	c1(ccccc1C(=O)O)CC	3.77	5.283	1.513	4.960	1.190	4.829	1.059	3.589	-0.181	2.661	-1.109
11	c1ccc(cc1Cl)C(=O)O	3.83	5.289	1.459	5.295	1.465	4.758	0.928	3.661	-0.169	1.913	-1.917
12	c1ccc(cc1F)C(=O)O	3.87	5.409	1.539	5.385	1.515	4.838	0.968	3.711	-0.159	1.537	-2.333
13	c1ccc(c(c1)C)C(=O)O	3.91	5.435	1.525	5.211	1.301	4.945	1.035	4.026	0.116	2.390	-1.520
14	c1(ccc(cc1)C(=O)O)Cl	3.99	5.609	1.619	5.602	1.612	5.077	1.087	4.239	0.249	1.820	-2.170
15	c1ccc(cc1O)C(=O)O	4.08	6.161	2.081	6.117	2.037	5.576	1.496	4.527	0.447	2.649	-1.431
16	c1ccc(c(c1)OC)C(=O)O	4.09	5.384	1.294	6.427	2.337	4.749	0.659	4.071	-0.019	2.139	-1.951
17	c1cc(cc(c1)OC)C(=O)O	4.09	6.224	2.134	6.155	2.065	5.518	1.428	4.365	0.275	2.419	-1.671
18	c1(ccc(cc1)C(=O)O)F	4.14	5.915	1.775	5.820	1.680	5.304	1.164	4.355	0.215	2.142	-1.998
19	c1(cccc(c1)C(=O)O)OCC	4.17	6.159	1.989	5.811	1.641	5.595	1.425	4.967	0.797	2.348	-1.822
20	c1(ccccc1C(=O)O)OCC	4.21	5.388	1.178	5.127	0.917	5.104	0.894	3.863	-0.347	2.123	-2.087
21	c1c(cc(cc1)C(=O)O)C	4.24	6.181	1.941	6.003	1.763	5.692	1.452	4.743	0.503	2.257	-1.983
22	c1(ccc(cc1)C(=O)O)C	4.34	6.406	2.066	5.786	1.446	5.569	1.229	4.816	0.476	2.299	-2.041
23	c1(ccc(cc1)C(=O)O)C(C)C	4.35	6.350	2.000	6.062	1.712	5.726	1.376	4.977	0.627	2.338	-2.012
24	c1(ccc(cc1)C(=O)O)CC	4.35	5.801	1.451	5.564	1.214	5.061	0.711	4.228	-0.122	2.350	-2.000
25	c1(ccc(cc1)C(=O)O)OCC	4.45	6.632	2.182	6.275	1.825	5.863	1.413	4.648	0.198	2.828	-1.622
26	c1c(ccc(c1)OC)C(=O)O	4.47	6.535	2.065	6.349	1.879	5.792	1.322	4.809	0.339	2.986	-1.484
27	c1cc(ccc1O)C(=O)O	4.58	6.534	1.954	6.093	1.513	5.719	1.139	4.931	0.351	3.104	-1.476
28	c1(ccc(cc1)C(=O)O)N(=O)=O	4.92	4.672	-0.248	4.885	-0.035	4.351	-0.569	3.548	-1.372	2.467	-2.453
	MAE			1.941		1.445		1.169		0.350		1.448
	MSE			1.723		1.526		1.217		0.459		1.603
	MaxAbsError			2.441		2.337		1.997		1.372		2.453

Table D15: Predictions on carbon acids from the Thapa & Raghavachari **Set-I** dataset[199]. QupKake was not trained on carbon acids, which resulted in no predictions.

			SMD only		SMD + 1 water		SMD + 2 waters		SMD + 3 waters		QupKake	
S.N.	SMILES	pK_a^{Exp}	pK_a	ΔpK_a	pK_a	ΔpK_a	pK_a	ΔpK_a	pK_a	ΔpK_a	pK_a	ΔpK_a
1	c1ccc2c(c1)c1c([C@H]2C(=O)C)cccc1	9.9	13.848	3.948	12.932	3.032	11.461	1.561	9.817	-0.083	–	–
2	C(=O)[C@H](c1ccccc1)c1ccccc1	10.4	14.953	4.553	14.159	3.759	12.715	2.315	10.325	-0.074	–	–
3	c1ccc2c(c1)c1c([C@H]2C(=O)SC)cccc1	10.5	13.372	2.872	12.973	2.473	11.621	1.121	10.164	-0.336	–	–
4	c1ccc2c(c1)c1c([C@H]2C(=O)OC)cccc1	11.5	13.812	2.312	12.647	1.147	12.602	1.102	10.845	-0.665	–	–
5	C(=O)Cc1ccccc1	13.1	18.727	5.627	18.632	5.532	14.54	1.44	13.204	0.104	–	–
6	C(=O)(CNC(=O)C)c1ccc(cc1)C	14.8	20.645	5.845	18.196	3.396	–	–	15.005	0.205	–	–
7	C(=O)C	16.7	23.613	6.913	21.613	4.913	19.534	2.834	17.323	0.623	–	–
8	C(=O)(C)c1ccccc1	18.3	24.955	6.655	22.903	4.603	20.931	2.631	18.821	0.521	–	–
9	C(=O)(C)c1ccc(cc1)C	19.2	25.511	6.311	23.53	4.33	21.274	2.075	20.083	0.883	–	–
10	C(=O)(C)C	19.3	26.728	7.428	24.793	5.493	22.933	3.633	21.023	1.723	–	–
	MAE			5.250		3.870		2.079		0.521		–
	MSE			5.250		3.870		2.079		0.291		–
	MaxAbsError			7.430		5.530		3.633		1.723		–

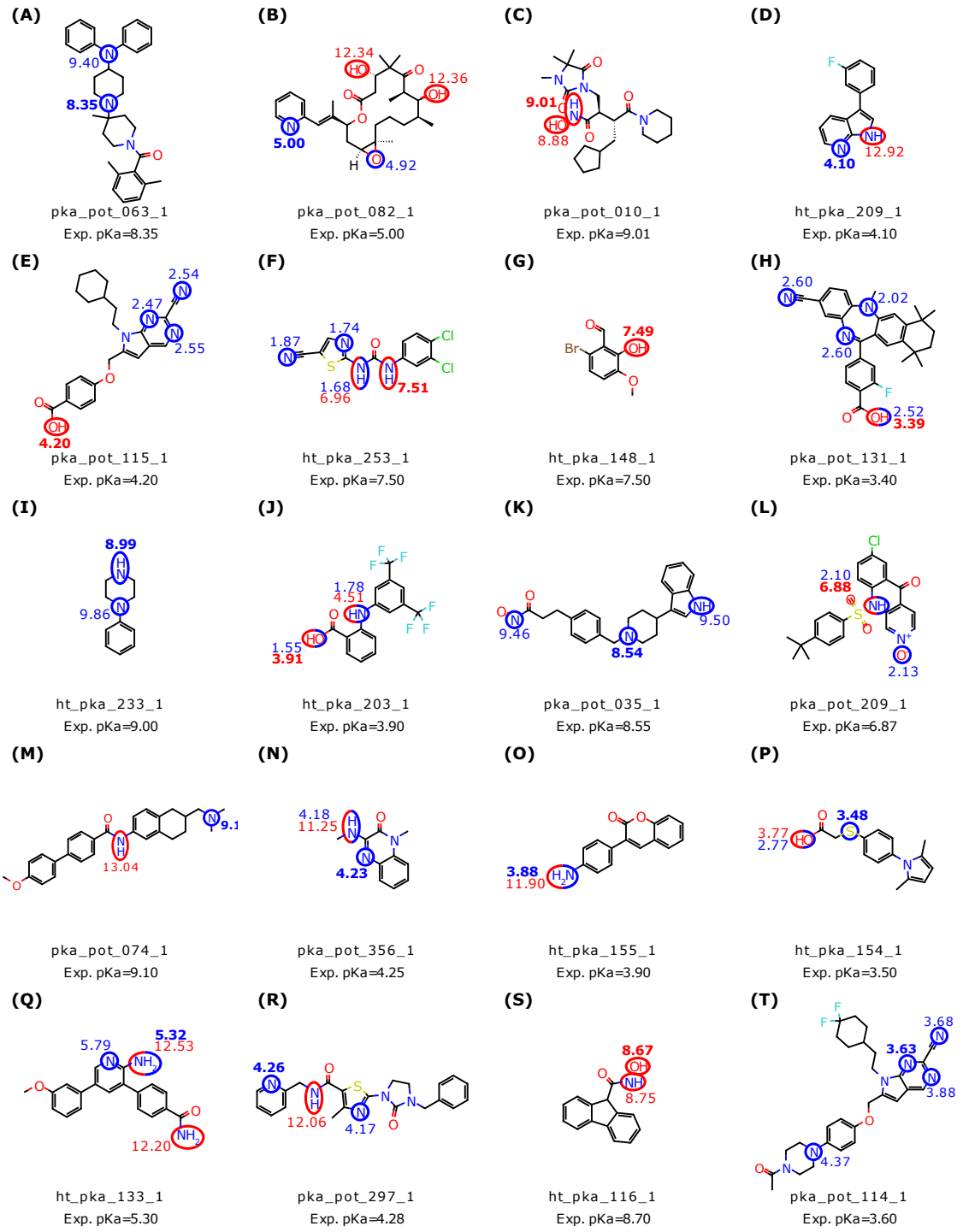


Figure D19: The top 20 molecules from the Novartis test set with the **most** accurate micro-pK_a prediction. The acidic and basic micro-pK_a values, as well as the atom they belongs to, are shown in red and blue, respectively. The micro-pK_a that is closest to the experimental value is shown in bold. The name, as it appear in the dataset, and the experimental pK_a value of each molecule are shown under the molecule.

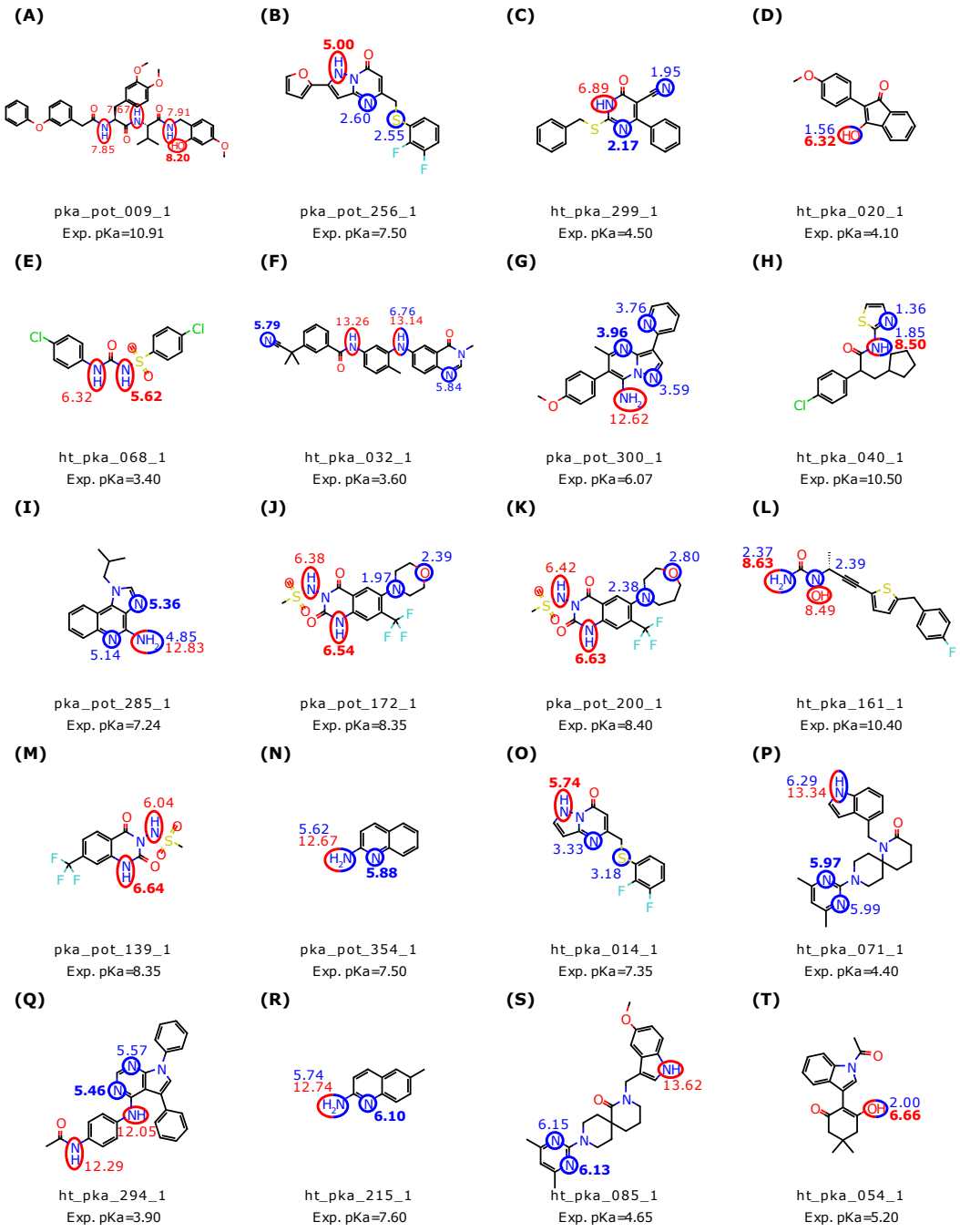
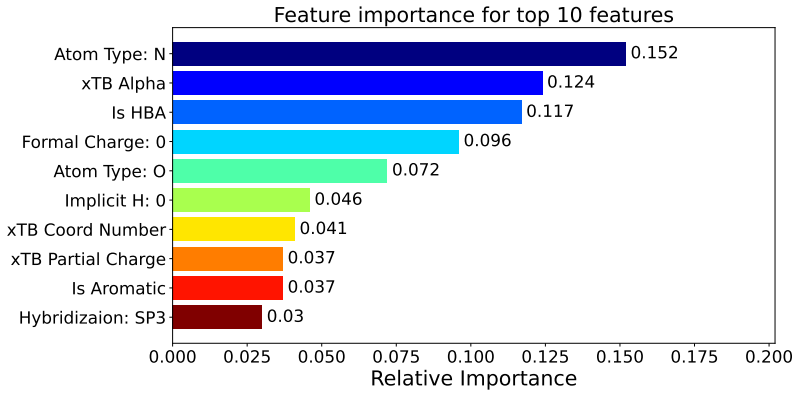
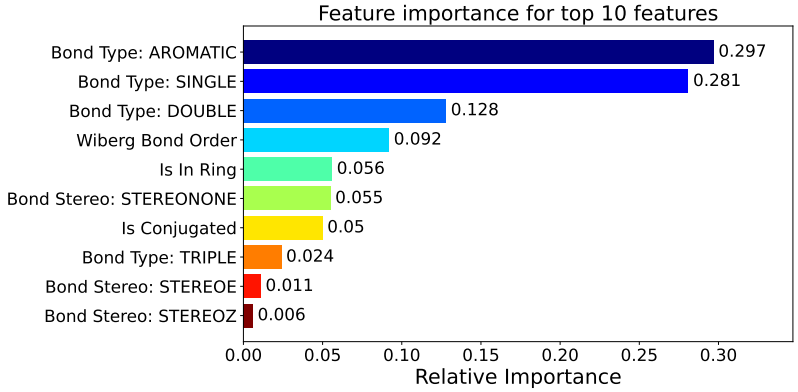


Figure D20: The bottom 20 molecules from the Novartis test set with the **least** accurate micro-pK_a prediction. The acidic and basic micro-pK_a values, as well as the atom they belongs to, are shown in red and blue, respectively. The micro-pK_a that is closest to the experimental value is shown in bold. The name, as it appear in the dataset, and the experimental pK_a value of each molecule are shown under the molecule.

(a)



(b)



(c)

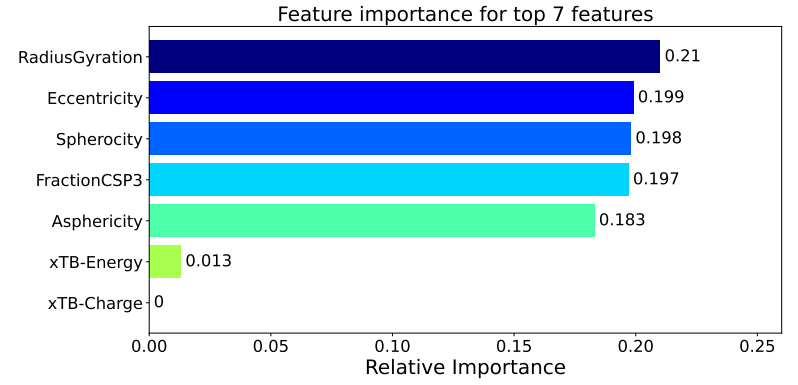


Figure D21: The normalized absolute relative importance of the a) atomic features, b) bond features and c) molecular features for the micro-pK_a prediction model.

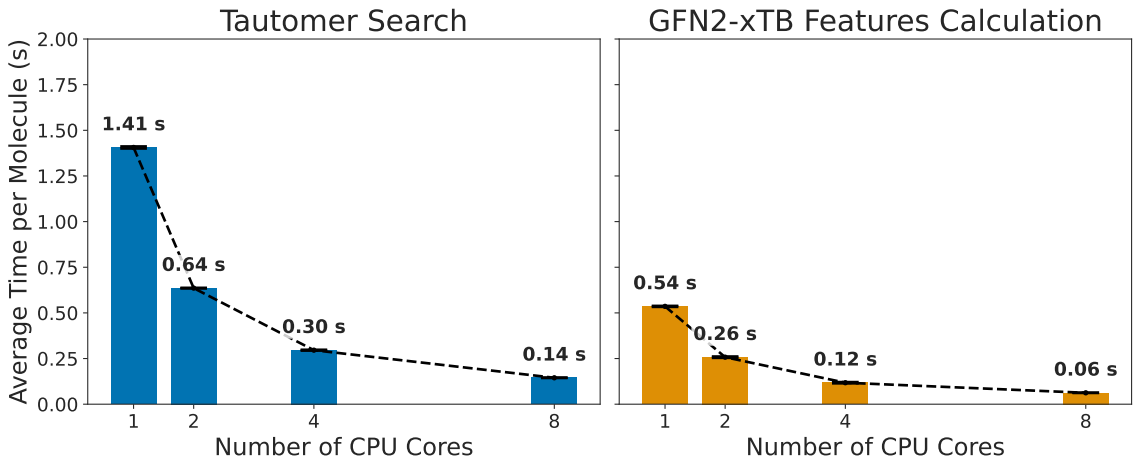


Figure D22: Tautomer search and GFN2-xTB features calculations average compute time per molecule across the 280 molecules in the Novartis test set as a function of the number of CPU cores.

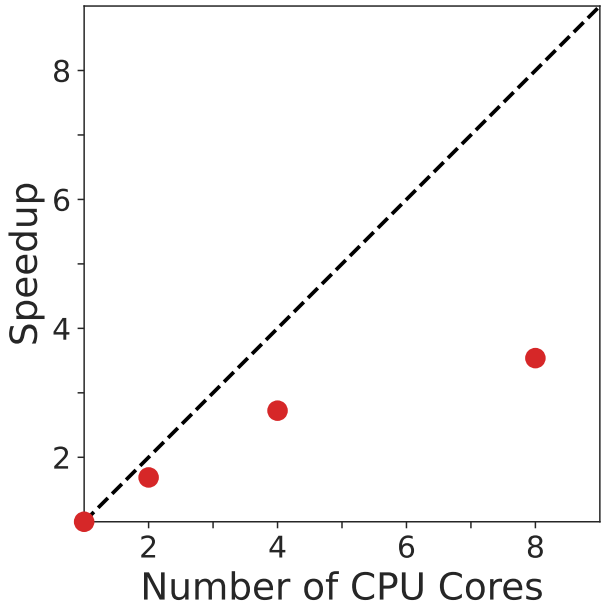


Figure D23: Compute time speedup rate across the 280 molecules in the Novartis test set as a function of CPU cores.

Bibliography

[1] M. Işık, A. S. Rustenburg, A. Rizzi, M. R. Gunner, D. L. Mobley, and J. D. Chodera, "Overview of the SAMPL6 pKa challenge: Evaluating small molecule microscopic and macroscopic pKa predictions," *Journal of Computer-Aided Molecular Design*, vol. 35, pp. 131–166, Feb. 2021.

[2] "The SAMPL6 Blind Prediction Challenges for Computational Chemistry." The SAMPL Challenges, Sept. 2023.

[3] "The SAMPL7 Blind Prediction Challenges for Computational Chemistry." https://github.com/samplchallenges/SAMPL7.

[4] "The SAMPL8 Blind Prediction Challenges for Computational Chemistry." The SAMPL Challenges, May 2023.

[5] R. C. Johnston, K. Yao, Z. Kaplan, M. Chelliah, K. Leswing, S. Seekins, S. Watts, D. Calkins, J. Chief Elk, S. V. Jerome, M. P. Repasky, and J. C. Shelley, "Epik: pKa and Protonation State Prediction through Machine Learning," *Journal of Chemical Theory and Computation*, vol. 19, pp. 2380–2388, Apr. 2023.

[6] F. Mayr, M. Wieder, O. Wieder, and T. Langer, "Improving Small Molecule pKa Prediction Using Transfer Learning With Graph Neural Networks," *Frontiers in Chemistry*, vol. 10, p. 866585, 2022.

[7] Y. Kim, S. A. Choulis, J. Nelson, D. D. C. Bradley, S. Cook, and J. R. Durrant, "Composition and annealing effects in polythiophene/fullerene solar cells," *Journal of Materials Science*, vol. 40, no. 6, pp. 1371–1376, 2005.

[8] M. Zhang, X. Guo, W. Ma, H. Ade, and J. Hou, "A polythiophene derivative with superior properties for practical application in polymer solar cells," *Advanced Materials*, vol. 26, pp. 5880–5885, sep 2014.

[9] Z. G. Zhang, S. Zhang, J. Min, C. Cui, H. Geng, Z. Shuai, and Y. Li, "Side chain engineering of polythiophene derivatives with a thienylene-vinylene conjugated side chain for application in polymer solar cells," *Macromolecules*, vol. 45, pp. 2312–2320, mar 2012.

[10] F. Wang, H. Gu, and T. M. Swager, "Carbon nanotube/polythiophene chemiresistive sensors for chemical warfare agents," *Journal of the American Chemical Society*, vol. 130, pp. 5392–5393, apr 2008.

[11] P. Schottland, M. Bouguettaya, and C. Chevrot, "Soluble polythiophene derivatives for NO2 sensing applications," *Synthetic Metals*, vol. 102, p. 1325, jun 1999.

[12] L. Wang, Q. Feng, X. Wang, M. Pei, and G. Zhang, "A novel polythiophene derivative as a sensitive colorimetric and fluorescent sensor for anionic surfactants in water," *New Journal of Chemistry*, vol. 36, pp. 1897–1901, aug 2012.

[13] B. H. Barboza, O. P. Gomes, and A. Batagin-Neto, "Polythiophene derivatives as chemical sensors: a DFT study on the influence of side groups," *Journal of Molecular Modeling*, vol. 27, p. 17, jan 2021.

[14] S. K. Kang, J. H. Kim, J. An, E. K. Lee, J. Cha, G. Lim, Y. S. Park, and D. J. Chung, "Synthesis of polythiophene derivatives and their application for electrochemical DNA sensor," *Polymer Journal*, vol. 36, pp. 937–942, jan 2004.

[15] A. L. Ding, J. Pei, Y. H. Lai, and W. Huang, "Phenylene-functionalized polythiophene derivatives for light-emitting diodes: Their synthesis, characterization and properties," *Journal of Materials Chemistry*, vol. 11, pp. 3082–3086, nov 2001.

[16] A. Facchetti, "π-Conjugated Polymers for Organic Electronics and Photovoltaic Cell Applications†," *Chemistry of Materials*, vol. 23, pp. 733–758, feb 2010.

[17] T. Tadesse, "Application of Conjugated Organic Polymers for Photovoltaic's: Review," 2018.

[18] D. T. Manallack, "The pKa Distribution of Drugs: Application to Drug Discovery," *Perspectives in Medicinal Chemistry*, vol. 1, p. 1177391X0700100, Jan. 2007.

[19] D. T. Manallack, R. J. Prankerd, E. Yuriev, T. I. Oprea, and D. K. Chalmers, "The significance of acid/base properties in drug discovery," *Chemical Society Reviews*, vol. 42, pp. 485–496, Dec. 2012.

[20] L. Gaohua, X. Miao, and L. Dou, "Crosstalk of physiological pH and chemical pKa under the umbrella of physiologically based pharmacokinetic modeling of drug absorption, distribution, metabolism, excretion, and toxicity," *Expert Opinion on Drug Metabolism & Toxicology*, vol. 17, pp. 1103–1124, Sept. 2021.

[21] H. Sahu and H. Ma, "Unraveling Correlations between Molecular Properties and Device Parameters of Organic Solar Cells Using Machine Learning," *Journal of Physical Chemistry Letters*, vol. 10, pp. 7277–7284, nov 2019.

[22] M. Rinderle, W. Kaiser, A. Mattoni, and A. Gagliardi, "Machine-Learned Charge Transfer Integrals for Multiscale Simulations in Organic Thin Films," *Journal of Physical Chemistry C*, vol. 124, pp. 17733–17743, aug 2020.

[23] D. Padula, J. D. Simpson, and A. Troisi, "Combining electronic and structural features in machine learning models to predict organic solar cells properties," *Materials Horizons*, vol. 6, pp. 343–349, feb 2019.

[24] D. Padula and A. Troisi, "Concurrent Optimization of Organic Donor–Acceptor Pairs through Machine Learning," *Advanced Energy Materials*, vol. 9, p. 1902463, oct 2019.

[25] C. Chen, Y. Zuo, W. Ye, X. Li, Z. Deng, and S. P. Ong, "A Critical Review of Machine Learning of Energy Materials," *Advanced Energy Materials*, vol. 10, p. 1903242, feb 2020.

[26] T. Sato, T. Honma, and S. Yokoyama, "Combining machine learning and pharmacophore-based interaction fingerprint for in silico screening," *Journal of Chemical Information and Modeling*, vol. 50, pp. 170–185, jan 2010.

[27] J. Vamathevan, D. Clark, P. Czodrowski, I. Dunham, E. Ferran, G. Lee, B. Li, A. Madabhushi, P. Shah, M. Spitzer, and S. Zhao, "Applications of machine learning in drug discovery and development," jun 2019.

[28] M. Misra, D. Andrienko, B. Baumeier, J.-L. Faulon, and O. A. von Lilienfeld, "Toward quantitative structure–property relationships for charge transfer rates of polycyclic aromatic hydrocarbons," *Journal of Chemical Theory and Computation*, vol. 7, pp. 2549–2555, July 2011.

[29] S. Atahan-Evrenk and F. B. Atalay, "Prediction of Intramolecular Reorganization Energy Using Machine Learning," *Journal of Physical Chemistry A*, vol. 123, pp. 7855–7863, sep 2019.

[30] J. Wu, Y. Kang, P. Pan, and T. Hou, "Machine learning methods for pKa prediction of small molecules: Advances and challenges," *Drug Discovery Today*, vol. 27, p. 103372, Dec. 2022.

[31] J. Wu, Y. Wan, Z. Wu, S. Zhang, D. Cao, C.-Y. Hsieh, and T. Hou, "MF-SuP-pKa: Multi-fidelity modeling with subgraph pooling mechanism for pKa prediction," *Acta Pharmaceutica Sinica B*, vol. 13, pp. 2572–2584, June 2023.

[32] M. Baltruschat and P. Czodrowski, "Machine learning meets pKa," *F1000Research*, vol. 9, p. 113, Apr. 2020.

[33] X. Pan, H. Wang, C. Li, J. Z. H. Zhang, and C. Ji, "MolGpka: A Web Server for Small Molecule pKa Prediction Using a Graph-Convolutional Neural Network," *Journal of Chemical Information and Modeling*, vol. 61, pp. 3159–3165, July 2021.

[34] Z. Bao and A. J. Lovinger, "Soluble regioregular polythiophene derivatives as semiconducting materials for field-effect transistors," *Chemistry of Materials*, vol. 11, no. 9, pp. 2607–2612, 1999.

[35] R. Porrazzo, S. Bellani, A. Luzio, C. Bertarelli, G. Lanzani, M. Caironi, and M. R. Antognazza, "Field-effect and capacitive properties of water-gated transistors based on polythiophene derivatives," *APL Materials*, vol. 3, p. 014905, jan 2015.

[36] G. R. Hutchison, M. A. Ratner, and T. J. Marks, "Hopping transport in conductive heterocyclic oligomers: Reorganization energies and substituent effects," *Journal of the American Chemical Society*, vol. 127, pp. 2339–2350, feb 2005.

[37] J. Cornil, D. Beljonne, J. P. Calbert, and J. L. Brédas, "Interchain interactions in organic π-conjugated materials: Impact on electronic structure, optical response, and charge transport," jul 2001.

[38] S. S. Zade and M. Bendikov, "Study of Hopping Transport in Long Oligothiophenes and Oligoselenophenes: Dependence of Reorganization Energy on Chain Length," *Chemistry - A European Journal*, vol. 14, pp. 6734–6741, jul 2008.

[39] K. M. Kleinow and M. S. Goodrich, "Environmental Aquatic Toxicology," in *Basic Environmental Toxicology*, CRC Press, 1994.

[40] S. I. Kang and Y. H. Bae, "pH-Induced solubility transition of sulfonamide-based polymers," *Journal of Controlled Release*, vol. 80, pp. 145–155, Apr. 2002.

[41] T. Wayne Schultz, "The use of the ionization constant (pKa) in selecting models of toxicity in phenols," *Ecotoxicology and Environmental Safety*, vol. 14, pp. 178–183, Oct. 1987.

[42] E. M. Ryan, C. B. Breslin, S. E. Moulton, and G. G. Wallace, "The effect of dopant pKa and the solubility of corresponding acid on the electropolymerisation of pyrrole," *Electrochimica Acta*, vol. 92, pp. 276–284, Mar. 2013.

[43] J. Nilsson and L. Baltzer, "Reactive-Site Design in Folded-Polypeptide Catalysts-The Leaving Group pKa of Reactive Esters Sets the Stage for Cooperativity in Nucleophilic and General-Acid Catalysis," *Chemistry – A European Journal*, vol. 6, no. 12, pp. 2214–2220, 2000.

[44] E. C. Kisgeropoulos, V. S. Bharadwaj, D. W. Mulder, and P. W. King, "The contribution of proton-donor pka on reactivity profiles of [fefe]-hydrogenases," 2022.

[45] P. Pracht, R. Wilcken, A. Udvarhelyi, S. Rodde, S. Grimme, c. A. Udvarhelyi, S. Rodde, and S. Grimme, "High accuracy quantum-chemistry-based calculation and blind prediction of macroscopic pKa values in the context of the SAMPL6 challenge," *Journal of Computer-Aided Molecular Design*, vol. 32, pp. 1139–1149, Oct. 2018.

[46] P. Pracht and S. Grimme, "Efficient Quantum-Chemical Calculations of Acid Dissociation Constants from Free-Energy Relationships," *The Journal of Physical Chemistry A*, vol. 125, pp. 5681–5692, July 2021.

[47] C. D. Navo and G. Jiménez-Osés, "Computer Prediction of pKa Values in Small Molecules and Proteins," *ACS Medicinal Chemistry Letters*, vol. 12, pp. 1624–1628, Nov. 2021.

[48] J. H. Jensen, C. J. Swain, and L. Olsen, "Prediction of pKa Values for Druglike Molecules Using Semiempirical Quantum Chemical Methods," *Journal of Physical Chemistry A*, vol. 121, pp. 699–707, Jan. 2017.

[49] N. Goudarzi and M. Goodarzi, "Prediction of the acidic dissociation constant (pKa) of some organic compounds using linear and nonlinear QSPR methods," *Molecular Physics*, vol. 107, pp. 1495–1503, Jan. 2009.

[50] P. Hunt, L. Hosseini-Gerami, T. Chrien, J. Plante, D. J. Ponting, and M. Segall, "Predicting pKa Using a Combination of Semi-Empirical Quantum Mechanics and Radial Basis Function Methods," *Journal of Chemical Information and Modeling*, vol. 60, pp. 2989–2997, June 2020.

[51] A. Zunger, "Inverse design in search of materials with target functionalities," *Nature Reviews Chemistry*, vol. 2, pp. 1–16, Mar. 2018.

[52] J. Wang, Y. Wang, and Y. Chen, "Inverse Design of Materials by Machine Learning," *Materials*, vol. 15, p. 1811, Jan. 2022.

[53] D. Folmsbee and G. Hutchison, "Assessing conformer energies using electronic structure and machine learning methods," *International Journal of Quantum Chemistry*, vol. 121, July 2020.

[54] C. Bannwarth, S. Ehlert, and S. Grimme, "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*, vol. 15, no. 3, pp. 1652–1671, 2019.

[55] O. D. Abarbanel and G. R. Hutchison, "Machine learning to accelerate screening for Marcus reorganization energies," *Journal of Chemical Physics*, vol. 155, no. 5, p. 54106, 2021.

[56] M. Mitchell, *An Introduction to Genetic Algorithms*. Cambridge, MA, USA: MIT Press, 1998.

[57] O. D. Abarbanel and G. R. Hutchison, "Using genetic algorithms to discover novel ground-state triplet conjugated polymers," *Physical Chemistry Chemical Physics*, vol. 25, no. 16, pp. 11278–11285, 2023.

[58] B. Greenstein, D. Elsey, and G. Hutchison, "Best Practices for Using Genetic Algorithms in Molecular Discovery," Feb. 2023.

[59] C. Lee, W. Yang, and R. G. Parr, "Development of the Colle-Salvetti correlation-energy formula into a functional of the electron density," *Physical Review B*, vol. 37, pp. 785–789, jan 1988.

[60] A. D. Becke, "Density-functional thermochemistry. III. The role of exact exchange," *The Journal of Chemical Physics*, vol. 98, pp. 5648–5652, apr 1993.

[61] N. M. O'Boyle, M. Banck, C. A. James, C. Morley, T. Vandermeersch, and G. R. Hutchison, "Open Babel: An open chemical toolbox," *Journal of Cheminformatics*, vol. 3, no. 1, p. 33, 2011.

[62] N. Yoshikawa and G. R. Hutchison, "Fast, efficient fragment-based coordinate generation for open babel," *Journal of Cheminformatics*, vol. 11, Aug. 2019.

[63] V. A. Rassolov, J. A. Pople, M. A. Ratner, and T. L. Windus, "6-31G* basis set for atoms K through Zn," *Journal of Chemical Physics*, vol. 109, pp. 1223–1229, jul 1998.

[64] M. J. Frisch, G. W. Trucks, H. B. Schlegel, G. E. Scuseria, M. A. Robb, J. R. Cheeseman, G. Scalmani, V. Barone, B. Mennucci, G. A. Petersson, H. Nakatsuji, M. Caricato, X. Li, H. P. Hratchian, A. F. Izmaylov, J. Bloino, G. Zheng, J. L. Sonnenberg, M. Hada, M. Ehara, K. Toyota, R. Fukuda, J. Hasegawa, M. Ishida, T. Nakajima, Y. Honda, O. Kitao, H. Nakai, T. Vreven, J. A. Montgomery, J. E. Peralta, F. Ogliaro, M. Bearpark, J. J. Heyd, E. Brothers, K. N. Kudin, V. N. Staroverov, R. Kobayashi, J. Normand, K. Raghavachari, A. Rendell, J. C. Burant, S. S. Iyengar, J. Tomasi, M. Cossi, N. Rega, J. M. Millam, M. Klene, J. E. Knox, J. B. Cross, V. Bakken, C. Adamo, J. Jaramillo, R. Gomperts, R. E. Stratmann, O. Yazyev, A. J. Austin, R. Cammi, C. Pomelli, J. W. Ochterski, R. L. Martin, K. Morokuma, V. G. Zakrzewski, G. A. Voth, P. Salvador, J. J. Dannenberg, S. Dapprich, A. D. Daniels, Ö. Farkas, J. B. Foresman, J. V. Ortiz, J. Cioslowski, and D. J. Fox, "Gaussian 09 Revision A.2," 2009.

[65] 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*, vol. 12, pp. 2825–2830, 2011.

[66] F. Chollet *et al.*, "Keras." `https://keras.io`, 2015.

[67] M. Abadi, A. Agarwal, P. Barham, E. Brevdo, Z. Chen, C. Citro, G. S. Corrado, A. Davis, J. Dean, M. Devin, S. Ghemawat, I. Goodfellow, A. Harp, G. Irving, M. Isard, Y. Jia, R. Jozefowicz, L. Kaiser, M. Kudlur, J. Levenberg, D. Mané, R. Monga, S. Moore, D. Murray, C. Olah, M. Schuster, J. Shlens, B. Steiner, I. Sutskever, K. Talwar, P. Tucker, V. Vanhoucke, V. Vasudevan, F. Viégas, O. Vinyals, P. Warden, M. Wattenberg, M. Wicke, Y. Yu, and X. Zheng, "TensorFlow: Large-scale machine learning on heterogeneous systems," 2015. Software available from tensorflow.org.

[68] D. Rogers and M. Hahn, "Extended-connectivity fingerprints," *Journal of Chemical Information and Modeling*, vol. 50, pp. 742–754, Apr. 2010.

[69] "RDKit: Open-source cheminformatics." `http://www.rdkit.org`, 2020. [Online; accessed 1-Mar-2021].

[70] L. Breiman, "Random Forests," *Machine Learning*, vol. 45, no. 1, pp. 5–32, 2001.

[71] J. Ye, J.-H. Chow, J. Chen, and Z. Zheng, "Stochastic gradient boosted distributed decision trees," in *Proceedings of the 18th ACM Conference on Information and Knowledge Management*, CIKM '09, (New York, NY, USA), p. 2061–2064, Association for Computing Machinery, 2009.

[72] A. E. Hoerl and R. W. Kennard, "Ridge regression: Biased estimation for nonorthogonal problems," *Technometrics*, vol. 42, no. 1, pp. 80–86, 2000.

[73] V. Vovk, "Kernel ridge regression," in *Empirical inference*, pp. 105–116, Springer, 2013.

[74] W. S. McCulloch and W. Pitts, "A logical calculus of the ideas immanent in nervous activity," *The bulletin of mathematical biophysics*, vol. 5, no. 4, pp. 115–133, 1943.

[75] J. Bergstra, D. Yamins, and D. C. B. T. P. o. t. t. I. C. o. M. Learning, "Making a Science of Model Search: Hyperparameter Optimization in Hundreds of Dimensions for Vision Architectures," feb 2013.

[76] M. Pumperla, "Hyperas." `https://github.com/maxpumperla/hyperas`, 2020.

[77] J. T. Barron, "Continuously differentiable exponential linear units," 2017.

[78] D. Misra, "Echo." `https://github.com/digantamisra98/Echo`, 2020.

[79] D. P. Kingma and J. L. Ba, "Adam: A method for stochastic optimization," in *3rd International Conference on Learning Representations, ICLR 2015 - Conference Track Proceedings*, International Conference on Learning Representations, ICLR, dec 2015.

[80] X. Ji and L. Fang, "Quinoidal Conjugated Polymers with Open-Shell Characters," *Polymer Chemistry*, 2021.

[81] A. E. London, H. Chen, M. A. Sabuj, J. Tropp, M. Saghayezhian, N. Eedugurala, B. A. Zhang, Y. Liu, X. Gu, B. M. Wong, N. Rai, M. K. Bowman, and J. D. Azoulay, "A high-spin ground-state donor-acceptor conjugated polymer," *Science Advances*, vol. 5, p. eaav2336, may 2019.

[82] T. L. Dexter Tam, C. K. Ng, S. L. Lim, E. Yildirim, J. Ko, W. L. Leong, S. W. Yang, and J. Xu, "Proquinoidal-conjugated polymer as an effective strategy for the

enhancement of electrical conductivity and thermoelectric properties," *Chemistry of Materials*, vol. 31, pp. 8543–8550, oct 2019.

[83] T. L. D. Tam, G. Wu, S. W. Chien, S. F. V. Lim, S. W. Yang, and J. Xu, "High Spin Pro-Quinoid Benzo[1,2-c;4,5-c']bisthiadiazole Conjugated Polymers for High-Performance Solution-Processable Polymer Thermoelectrics," *ACS Materials Letters*, vol. 2, pp. 147–152, feb 2020.

[84] I. Y. Kanal, S. G. Owens, J. S. Bechtel, and G. R. Hutchison, "Efficient computational screening of organic polymer photovoltaics," *The Journal of Physical Chemistry Letters*, vol. 4, pp. 1613–1623, Apr. 2013.

[85] L. Chan, G. M. Morris, and G. R. Hutchison, "Understanding conformational entropy in small molecules," *Journal of Chemical Theory and Computation*, vol. 17, no. 4, pp. 2099–2106, 2021.

[86] J. T. Blaskovits, K.-H. Lin, R. Fabregat, I. Swiderska, H. Wu, and C. Corminboeuf, "Is a single conformer sufficient to describe the reorganization energy of amorphous organic transport materials?," *ChemRxiv*, 2021.

[87] Y. Liu, L. Hua, S. Yan, and Z. Ren, "Halogenated π-conjugated polymeric emitters with thermally activated delayed fluorescence for highly efficient polymer light emitting diodes," *Nano Energy*, vol. 73, p. 104800, jul 2020.

[88] Y. Liu, S. Yan, and Z. Ren, "π-Conjugated polymeric light emitting diodes with sky-blue emission by employing thermally activated delayed fluorescence mechanism," *Chemical Engineering Journal*, vol. 417, p. 128089, aug 2021.

[89] J. Rao, L. Yang, X. Li, L. Zhao, S. Wang, H. Tian, J. Ding, and L. Wang, "Sterically-Locked Donor–Acceptor Conjugated Polymers Showing Efficient Thermally Activated Delayed Fluorescence," *Angewandte Chemie International Edition*, vol. 60, pp. 9635–9641, apr 2021.

[90] H. Noda, H. Nakanotani, and C. Adachi, "Excited state engineering for efficient reverse intersystem crossing," *Science Advances*, vol. 4, jun 2018.

[91] Y. Morita, S. Nishida, T. Murata, M. Moriguchi, A. Ueda, M. Satoh, K. Arifuku, K. Sato, and T. Takui, "Organic tailored batteries materials using stable open-shell molecules with degenerate frontier orbitals," *Nature Materials 2011 10:12*, vol. 10, pp. 947–951, oct 2011.

[92] K. Wang, L. Huang, N. Eedugurala, S. Zhang, M. A. Sabuj, N. Rai, X. Gu, J. D. Azoulay, and T. N. Ng, "Wide Potential Window Supercapacitors Using Open-Shell Donor–Acceptor Conjugated Polymers with Stable N-Doped States," *Advanced Energy Materials*, vol. 9, p. 1902806, dec 2019.

[93] M. Nakano, "Open-Shell-Character-Based Molecular Design Principles: Applications to Nonlinear Optics and Singlet Fission," *Chemical Record*, vol. 17, pp. 27–62, jan 2017.

[94] Z. Sun, Q. Ye, C. Chi, and J. Wu, "Low band gap polycyclic hydrocarbons: from closed-shell near infrared dyes and semiconductors to open-shell radicals," *Chemical Society Reviews*, vol. 41, pp. 7857–7889, nov 2012.

[95] Y. Huang and E. Egap, "Open-shell organic semiconductors: an emerging class of materials with novel properties," *Polymer Journal 2018 50:8*, vol. 50, pp. 603–614, may 2018.

[96] L. Huang, N. Eedugurala, A. Benasco, S. Zhang, K. S. Mayer, D. J. Adams, B. Fowler, M. M. Lockart, M. Saghayezhian, H. Tahir, E. R. King, S. Morgan, M. K. Bowman, X. Gu, J. D. Azoulay, L. Huang, N. Eedugurala, A. Benasco, S. Zhang, K. S. Mayer, D. J. Adams, H. Tahir, E. R. King, S. Morgan, X. Gu, J. D. Azoulay, B. Fowler, M. M. Lockart, M. K. Bowman, and M. Saghayezhian, "Open-Shell Donor-Acceptor Conjugated Polymers with High Electrical Conductivity," 2020.

[97] Z. Wu, W. Yao, A. E. London, J. D. Azoulay, and T. N. Ng, "Temperature-Dependent Detectivity of Near-Infrared Organic Bulk Heterojunction Photodiodes," *ACS Applied Materials and Interfaces*, vol. 9, pp. 1654–1660, jan 2017.

[98] D. B. Sulas, A. E. London, L. Huang, L. Xu, Z. Wu, T. N. Ng, B. M. Wong, C. W. Schlenker, J. D. Azoulay, and M. Y. Sfeir, "Preferential Charge Generation at Aggregate Sites in Narrow Band Gap Infrared Photoresponsive Polymer Semiconductors," *Advanced Optical Materials*, vol. 6, p. 1701138, apr 2018.

[99] Y. Joo, L. Huang, N. Eedugurala, A. E. London, A. Kumar, B. M. Wong, B. W. Boudouris, and J. D. Azoulay, "Thermoelectric Performance of an Open-Shell Donor–Acceptor Conjugated Polymer Doped with a Radical-Containing Small Molecule," *Macromolecules*, vol. 51, pp. 3886–3894, may 2018.

[100] Z. Wu, Y. Zhai, H. Kim, J. D. Azoulay, and T. N. Ng, "Emerging Design and Characterization Guidelines for Polymer-Based Infrared Photodetectors," *Accounts of Chemical Research*, vol. 51, pp. 3144–3153, dec 2018.

[101] S. Zhang, M. U. Ocheje, L. Huang, L. Galuska, Z. Cao, S. Luo, Y.-H. Cheng, D. Ehlenberg, R. B. Goodman, D. Zhou, Y. Liu, Y.-C. Chiu, J. D. Azoulay, S. Rondeau-Gagné, and X. Gu, "The Critical Role of Electron-Donating Thiophene Groups on the Mechanical and Thermal Properties of Donor–Acceptor Semiconducting Polymers," *Advanced Electronic Materials*, vol. 5, p. 1800899, may 2019.

[102] A. M. Asaduzzaman, K. Schmidt-D'Aloisio, Y. Dong, and M. Springborg, "Properties of polythiophene and related conjugated polymers: a density-functional study," *Phys. Chem. Chem. Phys.*, vol. 7, pp. 2714–2722, 2005.

[103] J. Huang, S. Lu, P.-A. Chen, K. Wang, Y. Hu, Y. Liang, M. Wang, and E. Reichmanis, "Rational design of a narrow-bandgap conjugated polymer using the quinoidal thieno[3,2-b]thiophene-based building block for organic field-effect transistor applications," *Macromolecules*, vol. 52, no. 12, pp. 4749–4756, 2019.

[104] L. Salem and C. Rowland, "The Electronic Properties of Diradicals," *Angewandte Chemie International Edition in English*, vol. 11, pp. 92–111, feb 1972.

[105] T. Y. Gopalakrishna, W. Zeng, X. Lu, and J. Wu, "From open-shell singlet diradicaloids to polyradicaloids," *Chemical Communications*, vol. 54, pp. 2186–2199, feb 2018.

[106] D. Bokhan and R. J. Bartlett, "Ab initio density functional theory for spin-polarized systems," *Chemical Physics Letters*, vol. 427, pp. 466–471, aug 2006.

[107] K. Hirao, "Multireference Møller—Plesset method," *Chemical Physics Letters*, vol. 190, pp. 374–380, mar 1992.

[108] T. Yanai, D. P. Tew, and N. C. Handy, "A new hybrid exchange-correlation functional using the Coulomb-attenuating method (CAM-B3LYP),"

[109] T. P. Kaloni, G. Schreckenbach, and M. S. Freund, "Band gap modulation in polythiophene and polypyrrole-based systems," *Scientific Reports*, vol. 6, no. 1, p. 36554, 2016.

[110] S. M. Bouzzine, G. Salgado-Morán, M. Hamidi, M. Bouachrine, A. G. Pacheco, and D. Glossman-Mitnik, "DFT Study of Polythiophene Energy Band Gap and Substitution Effects," 2015.

[111] JOHNSON III and D. RUSSELL, "NIST Computational Chemistry Comparison and Benchmark Database," 2020.

[112] K. Yamamoto, Y. Ie, M. Nitani, N. Tohnai, F. Kakiuchi, K. Zhang, W. Pisula, K. Asadi, P. W. M. Blom, and Y. Aso, "Oligothiophene quinoids containing a benzo[c]thiophene unit for the stabilization of the quinoidal electronic structure," *J. Mater. Chem. C*, vol. 6, no. 28, pp. 7493–7500, 2018.

[113] S. K. Singh, X. Crispin, and I. V. Zozoulenko, "Oxygen Reduction Reaction in Conducting Polymer PEDOT: Density Functional Theory Study," *The Journal of Physical Chemistry C*, vol. 121, no. 22, pp. 12270–12277, 2017.

[114] S. S. Zade and M. Bendikov, "From oligomers to polymer: Convergence in the homo-lumo gaps of conjugated oligomers," *Organic Letters*, vol. 8, no. 23, pp. 5243–5246, 2006.

[115] D. Weininger, "SMILES, a Chemical Language and Information System: 1: Introduction to Methodology and Encoding Rules," *Journal of Chemical Information and Computer Sciences*, vol. 28, pp. 31–36, feb 1988.

[116] D. Weininger, A. Weininger, and J. L. Weininger, "SMILES. 2. Algorithm for Generation of Unique SMILES Notation," *Journal of Chemical Information and Computer Sciences*, vol. 29, no. 2, pp. 97–101, 1989.

[117] D. Weininger, "Smiles. 3. Depict. Graphical Depiction of Chemical Structures," *Journal of Chemical Information and Computer Sciences*, vol. 30, pp. 237–243, aug 1990.

[118] T. A. Halgren, "Merck molecular force field. I. Basis, form, scope, parameterization, and performance of MMFF94," *Journal of Computational Chemistry*, vol. 17, no. 5-6, pp. 490–519, 1996.

[119] A. K. Rappe, C. J. Casewit, K. S. Colwell, W. A. G. III, and W. M. Skiff, "UFF, a full periodic table force field for molecular mechanics and molecular dynamics simulations," *Journal of the American Chemical Society*, vol. 114, pp. 10024–10035, dec 2002.

[120] J. G. Brandenburg, C. Bannwarth, A. Hansen, and S. Grimme, "B97-3c: A revised low-cost variant of the B97-D density functional method," *The Journal of Chemical Physics*, vol. 148, p. 064104, feb 2018.

[121] F. Neese, "The ORCA program system," *Wiley Interdisciplinary Reviews: Computational Molecular Science*, vol. 2, pp. 73–78, jan 2012.

[122] F. Neese, "Software update: the ORCA program system, version 4.0," *Wiley Interdisciplinary Reviews: Computational Molecular Science*, vol. 8, p. e1327, jan 2018.

[123] "Systematic optimization of long-range corrected hybrid density functionals," *The Journal of Chemical Physics*, vol. 128, p. 084106, feb 2008.

[124] F. Weigend, "Accurate Coulomb-fitting basis sets for H to Rn," *Physical Chemistry Chemical Physics*, vol. 8, pp. 1057–1065, feb 2006.

[125] N. M. O'boyle, A. L. Tenderholt, and K. M. Langner, "cclib: A library for package-independent computational chemistry algorithms," *Journal of Computational Chemistry*, vol. 29, pp. 839–845, apr 2008.

[126] O. D. Abarbanel, J. Rozon, and G. R. Hutchison, "Strategies for Computer-Aided Discovery of Novel Open-Shell Polymers," *Journal of Physical Chemistry Letters*, vol. 13, pp. 2158–2164, mar 2022.

[127] Y. Liu, T. Zhao, W. Ju, S. Shi, S. Shi, and S. Shi, "Materials discovery and design using machine learning," *Journal of Materiomics*, vol. 3, pp. 159–177, sep 2017.

[128] K. Guo, Z. Yang, C. H. Yu, and M. J. Buehler, "Artificial intelligence and machine learning in design of mechanical materials," *Materials Horizons*, vol. 8, pp. 1153–1172, apr 2021.

[129] R. Vasudevan, G. Pilania, and P. V. Balachandran, "Machine learning for materials design and discovery," *Journal of Applied Physics*, vol. 129, p. 070401, feb 2021.

[130] C. Darwin, *On the Origin of Species by Means of Natural Selection*. London: Murray, 1859. or the Preservation of Favored Races in the Struggle for Life.

[131] D. C. Hiener and G. R. Hutchison, "Pareto Optimization of Oligomer Polarizability and Dipole Moment Using a Genetic Algorithm," *The Journal of Physical Chemistry A*, vol. 126, pp. 2750–2760, may 2022.

[132] G. Van Rossum and F. L. Drake, *Python 3 Reference Manual*. Scotts Valley, CA: CreateSpace, 2009.

[133] B. L. Greenstein and G. R. Hutchison, "Organic Photovoltaic Efficiency Predictor: Data-Driven Models for Non-Fullerene Acceptor Organic Solar Cells," *The Journal of Physical Chemistry Letters*, vol. 13, pp. 4235–4243, may 2022.

[134] B. N. Norris, S. Zhang, C. M. Campbell, J. T. Auletta, P. Calvo-Marzal, G. R. Hutchison, and T. Y. Meyer, "Sequence matters: Modulating electronic and optical properties of conjugated oligomers via tailored sequence," *Macromolecules*, vol. 46, pp. 1384–1392, Feb. 2013.

[135] I. Y. Kanal, J. S. Bechtel, and G. R. Hutchison, "Sequence matters: Determining the sequence effect of electronic structure properties in p-conjugated polymers," in *ACS Symposium Series*, pp. 379–393, American Chemical Society, Jan. 2014.

[136] S. Zhang, G. R. Hutchison, and T. Y. Meyer, "Sequence effects in conjugated donor-acceptor trimers and polymers," *Macromolecular Rapid Communications*, vol. 37, pp. 882–887, Apr. 2016.

[137] S. Zhang, N. E. Bauer, I. Y. Kanal, W. You, G. R. Hutchison, and T. Y. Meyer, "Sequence effects in donor–acceptor oligomeric semiconductors comprising benzothiadiazole and phenylenevinylene monomers," *Macromolecules*, vol. 50, pp. 151–161, Dec. 2016.

[138] Y. S. Lin, G. D. Li, S. P. Mao, and J. D. Chai, "Long-range corrected hybrid density functionals with improved dispersion corrections," *Journal of Chemical Theory and Computation*, vol. 9, pp. 263–272, jan 2013.

[139] T. Yanai, D. P. Tew, and N. C. Handy, "A new hybrid exchange–correlation functional using the Coulomb-attenuating method (CAM-B3LYP)," *Chemical Physics Letters*, vol. 393, pp. 51–57, jul 2004.

[140] N. E. Jackson, B. M. Savoie, K. L. Kohlstedt, T. J. Marks, L. X. Chen, and M. A. Ratner, "Structural and Conformational Dispersion in the Rational Design of Conjugated Polymers," *Macromolecules*, vol. 47, pp. 987–992, Feb. 2014.

[141] L. Wilbraham, E. Berardo, L. Turcani, K. E. Jelfs, and M. A. Zwijnenburg, "High-Throughput Screening Approach for the Optoelectronic Properties of Conjugated Polymers," *Journal of Chemical Information and Modeling*, vol. 58, pp. 2450–2459, Dec. 2018.

[142] B. M. Savoie, N. E. Jackson, L. X. Chen, T. J. Marks, and M. A. Ratner, "Mesoscopic Features of Charge Generation in Organic Semiconductors," *Accounts of Chemical Research*, vol. 47, pp. 3385–3394, Nov. 2014.

[143] M. D. Hanwell, D. E. Curtis, D. C. Lonie, T. Vandermeersch, E. Zurek, and G. R. Hutchison, "Avogadro: an advanced semantic chemical editor, visualization, and analysis platform," *Journal of Cheminformatics*, vol. 4, p. 17, Aug. 2012.

[144] S. A. Wildman and G. M. Crippen, "Prediction of Physicochemical Parameters by Atomic Contributions," *Journal of Chemical Information and Computer Sciences*, vol. 39, pp. 868–873, Sept. 1999.

[145] B. M. Wong and J. G. Cordaro, "Electronic properties of vinylene-linked heterocyclic conducting polymers: Predictive design and rational guidance from DFT calculations," *Journal of Physical Chemistry C*, vol. 115, pp. 18333–18341, sep 2011.

[146] K. Kobayashi, M. S. Mohamed Ahmed, and A. Mori, "Introduction of ethynylene and thienylene spacers into 2,5-diarylthiazole and 2,5-diarylthiophene," *Tetrahedron*, vol. 62, pp. 9548–9553, oct 2006.

[147] W. I. Hung, Y. Y. Liao, C. Y. Hsu, H. H. Chou, T. H. Lee, W. S. Kao, and J. T. Lin, "High-Performance Dye-Sensitized Solar Cells Based on Phenothiazine Dyes Containing Double Anchors and Thiophene Spacers," *Chemistry – An Asian Journal*, vol. 9, pp. 357–366, jan 2014.

[148] M. Paramasivam, A. Gupta, A. M. Raynor, S. V. Bhosale, K. Bhanuprakash, and V. Jayathirtha Rao, "Small band gap D-π-A-π-D benzothiadiazole derivatives with low-lying HOMO levels as potential donors for applications in organic photovoltaics: a combined experimental and theoretical investigation," *RSC Advances*, vol. 4, pp. 35318–35331, aug 2014.

[149] A. Shuto, T. Kushida, T. Fukushima, H. Kaji, and S. Yamaguchi, "π-Extended Planarized Triphenylboranes with Thiophene Spacers," *Organic Letters*, vol. 15, no. 24, pp. 6234–6237, 2013.

[150] R. Rausch, D. Schmidt, D. Bialas, I. Krummenacher, H. Braunschweig, and F. Würthner, "Stable Organic (Bi)Radicals by Delocalization of Spin Density into the Electron-Poor Chromophore Core of Isoindigo," *Chemistry - A European Journal*, vol. 24, pp. 3420–3424, mar 2018.

[151] A. Rajca, "Organic Diradicals and Polyradicals: From Spin Coupling to Magnetism?," *Chemical Reviews*, vol. 94, pp. 871–893, jun 2002.

[152] M.-H. Lin, J.-F. Tsai, and C.-S. Yu, "A Review of Deterministic Optimization Methods in Engineering and Management," *Mathematical Problems in Engineering*, vol. 2012, p. 756023, June 2012.

[153] R. S. Bohacek, C. McMartin, and W. C. Guida, "The art and practice of structure-based drug design: A molecular modeling perspective," *Medicinal Research Reviews*, vol. 16, no. 1, pp. 3–50, 1996.

[154] K. Weiss, T. M. Khoshgoftaar, and D. Wang, "A survey of transfer learning," *Journal of Big Data*, vol. 3, p. 9, May 2016.

[155] Y. C. Martin, "Let's not forget tautomers," *Journal of Computer-Aided Molecular Design*, vol. 23, pp. 693–704, Oct. 2009.

[156] C. M. Baker, N. J. Kidley, K. Papachristos, M. Hotson, R. Carson, D. Gravestock, M. Pouliot, J. Harrison, and A. Dowling, "Tautomer Standardization in Chemical Databases: Deriving Business Rules from Quantum Chemistry," *Journal of Chemical Information and Modeling*, vol. 60, pp. 3781–3791, Aug. 2020.

[157] D. K. Dhaked, W.-D. Ihlenfeldt, H. Patel, V. Delannée, and M. C. Nicklaus, "Toward a Comprehensive Treatment of Tautomerism in Chemoinformatics Including in InChI V2," *Journal of Chemical Information and Modeling*, vol. 60, pp. 1253–1275, Mar. 2020.

[158] S. Ehlert, M. Stahn, S. Spicher, and S. Grimme, "Robust and Efficient Implicit Solvation Model for Fast Semiempirical Methods," *Journal of Chemical Theory and Computation*, vol. 17, pp. 4250–4261, July 2021.

[159] "Daylight Theory: SMARTS - A Language for Describing Molecular Patterns." https://www.daylight.com/dayhtml/doc/theory/theory.smarts.html.

[160] P. Pracht, F. Bohle, and S. Grimme, "Automated exploration of the low-energy chemical space with fast quantum chemical methods," *Physical Chemistry Chemical Physics*, vol. 22, pp. 7169–7192, Apr. 2020.

[161] P. Pracht, C. A. Bauer, and S. Grimme, "Automated and efficient quantum chemical determination and energetic ranking of molecular protonation sites," *Journal of Computational Chemistry*, vol. 38, no. 30, pp. 2618–2631, 2017.

[162] K. Riedmiller, P. Reiser, E. Bobkova, K. Maltsev, G. Gryn'ova, P. Friederich, and F. Gräter, "Substituting density functional theory in reaction barrier calculations for hydrogen atom transfer in proteins," *Chemical Science*, vol. 15, pp. 2518–2527, Feb. 2024.

[163] M. Davies, M. Nowotka, G. Papadatos, N. Dedman, A. Gaulton, F. Atkinson, L. Bellis, and J. P. Overington, "ChEMBL web services: Streamlining access to drug discovery data and utilities," *Nucleic Acids Research*, vol. 43, pp. W612–W620, July 2015.

[164] D. Mendez, A. Gaulton, A. P. Bento, J. Chambers, M. De Veij, E. Félix, M. P. Magariños, J. F. Mosquera, P. Mutowo, M. Nowotka, M. Gordillo-Marañón, F. Hunter, L. Junco, G. Mugumbate, M. Rodriguez-Lopez, F. Atkinson, N. Bosc, C. J. Radoux, A. Segura-Cabrera, A. Hersey, and A. R. Leach, "ChEMBL: Towards direct deposition of bioassay data," *Nucleic Acids Research*, vol. 47, pp. D930–D940, Jan. 2019.

[165] ChemAxon Marvin Suite, ChemAxon Inc. Available from: http://www.chemaxon.com.

[166] N. M. O'Boyle, M. Banck, C. A. James, C. Morley, T. Vandermeersch, and G. R. Hutchison, "Open Babel: An open chemical toolbox," *Journal of Cheminformatics*, vol. 3, p. 33, Oct. 2011.

[167] O. D. Abarbanel and G. R. Hutchison, "Machine learning to accelerate screening for Marcus reorganization energies," *Journal of Chemical Physics*, vol. 155, p. 54106, Aug. 2021.

[168] G. A. Pinheiro, J. Mucelini, M. D. Soares, R. C. Prati, J. L. F. Da Silva, and M. G. Quiles, "Machine Learning Prediction of Nine Molecular Properties Based on the SMILES Representation of the QM9 Quantum-Chemistry Dataset," *The Journal of Physical Chemistry A*, vol. 124, pp. 9854–9866, Nov. 2020.

[169] J. Vamathevan, D. Clark, P. Czodrowski, I. Dunham, E. Ferran, G. Lee, B. Li, A. Madabhushi, P. Shah, M. Spitzer, and S. Zhao, "Applications of machine learning in drug discovery and development," *Nature Reviews Drug Discovery*, vol. 18, pp. 463–477, June 2019.

[170] Z. Wu, B. Ramsundar, E. N. Feinberg, J. Gomes, C. Geniesse, A. S. Pappu, K. Leswing, and V. Pande, "MoleculeNet: A benchmark for molecular machine learning," *Chemical Science*, vol. 9, pp. 513–530, Jan. 2018.

[171] K. Gubaev, E. V. Podryabinkin, and A. V. Shapeev, "Machine learning of molecular properties: Locality and active learning," *The Journal of Chemical Physics*, vol. 148, p. 241727, Apr. 2018.

[172] N. Fedik, R. Zubatyuk, M. Kulichenko, N. Lubbers, J. S. Smith, B. Nebgen, R. Messerly, Y. W. Li, A. I. Boldyrev, K. Barros, O. Isayev, and S. Tretiak, "Extending machine learning beyond interatomic potentials for predicting molecular properties," *Nature Reviews Chemistry*, vol. 6, pp. 653–672, Sept. 2022.

[173] R. Ramakrishnan, P. O. Dral, M. Rupp, and O. A. von Lilienfeld, "Big Data Meets Quantum Chemistry Approximations: The Δ-Machine Learning Approach," *Journal of Chemical Theory and Computation*, vol. 11, pp. 2087–2096, May 2015.

[174] K. T. Schütt, F. Arbabzadah, S. Chmiela, K. R. Müller, and A. Tkatchenko, "Quantum-chemical insights from deep tensor neural networks," *Nature Communications*, vol. 8, p. 13890, Jan. 2017.

[175] F. A. Faber, L. Hutchison, B. Huang, J. Gilmer, S. S. Schoenholz, G. E. Dahl, O. Vinyals, S. Kearnes, P. F. Riley, and O. A. von Lilienfeld, "Prediction Errors of Molecular Machine Learning Models Lower than Hybrid DFT Error," *Journal of Chemical Theory and Computation*, vol. 13, pp. 5255–5264, Nov. 2017.

[176] J. Xiong, Z. Xiong, K. Chen, H. Jiang, and M. Zheng, "Graph neural networks for automated de novo drug design," *Drug Discovery Today*, vol. 26, pp. 1382–1393, June 2021.

[177] D. Jiang, Z. Wu, C.-Y. Hsieh, G. Chen, B. Liao, Z. Wang, C. Shen, D. Cao, J. Wu, and T. Hou, "Could graph neural networks learn better molecular representation for drug discovery? A comparison study of descriptor-based and graph-based models," *Journal of Cheminformatics*, vol. 13, p. 12, Feb. 2021.

[178] P. Reiser, M. Neubert, A. Eberhard, L. Torresi, C. Zhou, C. Shao, H. Metni, C. van Hoesel, H. Schopmans, T. Sommer, and P. Friederich, "Graph neural networks for materials science and chemistry," *Communications Materials*, vol. 3, pp. 1–18, Nov. 2022.

[179] V. Fung, J. Zhang, E. Juarez, and B. G. Sumpter, "Benchmarking graph neural networks for materials chemistry," *npj Computational Materials*, vol. 7, pp. 1–8, June 2021.

[180] O. Wieder, S. Kohlbacher, M. Kuenemann, A. Garon, P. Ducrot, T. Seidel, and T. Langer, "A compact review of molecular property prediction with graph neural networks," *Drug Discovery Today: Technologies*, vol. 37, pp. 1–12, Dec. 2020.

[181] Z. Yang, M. Chakraborty, and A. D. White, "Predicting chemical shifts with graph neural networks," *Chemical Science*, vol. 12, no. 32, pp. 10802–10809, 2021.

[182] M. Fey and J. E. Lenssen, "Fast graph representation learning with PyTorch Geometric," in *ICLR Workshop on Representation Learning on Graphs and Manifolds*, 2019.

[183] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga, A. Desmaison, A. Köpf, E. Yang, Z. DeVito, M. Raison, A. Tejani, S. Chilamkurthy, B. Steiner, L. Fang, J. Bai, and S. Chintala, "PyTorch: An imperative style, high-performance deep learning library," in *Proceedings of the 33rd International Conference on Neural Information Processing Systems*, no. 721, pp. 8026–8037, Red Hook, NY, USA: Curran Associates Inc., Dec. 2019.

[184] W. Falcon and The PyTorch Lightning team, "PyTorch Lightning," Mar. 2019.

[185] T. N. Kipf and M. Welling, "Semi-Supervised Classification with Graph Convolutional Networks," Feb. 2017.

[186] P. Veličković, G. Cucurull, A. Casanova, A. Romero, P. Liò, and Y. Bengio, "Graph Attention Networks," Feb. 2018.

[187] Y. Shi, Z. Huang, S. Feng, H. Zhong, W. Wang, and Y. Sun, "Masked Label Prediction: Unified Message Passing Model for Semi-Supervised Classification," May 2021.

[188] T. Akiba, S. Sano, T. Yanase, T. Ohta, and M. Koyama, "Optuna: A next-generation hyperparameter optimization framework," in *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2019.

[189] M. Sundararajan, A. Taly, and Q. Yan, "Axiomatic Attribution for Deep Networks," June 2017.

[190] N. Kokhlikyan, V. Miglani, M. Martin, E. Wang, B. Alsallakh, J. Reynolds, A. Melnikov, N. Kliushkina, C. Araya, S. Yan, and O. Reblitz-Richardson, "Captum: A unified and generic model interpretability library for PyTorch," Sept. 2020.

[191] Q. Yang, Y. Li, J.-D. Yang, Y. Liu, L. Zhang, S. Luo, and J.-P. Cheng, "http://ibond.nankai.edu.cn."

[192] G. FOODY, M. B. McCULLOCH, and W. B. YATES, "The effect of training set size and composition on artificial neural network classification," *International Journal of Remote Sensing*, vol. 16, pp. 1707–1723, June 1995.

[193] C. Cai, S. Wang, Y. Xu, W. Zhang, K. Tang, Q. Ouyang, L. Lai, and J. Pei, "Transfer Learning for Drug Discovery," *Journal of Medicinal Chemistry*, vol. 63, pp. 8683–8694, Aug. 2020.

[194] C. Tan, F. Sun, T. Kong, W. Zhang, C. Yang, and C. Liu, "A Survey on Deep Transfer Learning," in *Artificial Neural Networks and Machine Learning – ICANN 2018* (V. Kůrková, Y. Manolopoulos, B. Hammer, L. Iliadis, and I. Maglogiannis, eds.), Lecture Notes in Computer Science, (Cham), pp. 270–279, Springer International Publishing, 2018.

[195] F. Tsung, K. Zhang, L. Cheng, and Z. Song, "Statistical transfer learning: A review and some extensions to statistical process control," *Quality Engineering*, vol. 30, pp. 115–128, Jan. 2018.

[196] C. D. Stern, C. I. Bayly, D. G. A. Smith, J. Fass, L.-P. Wang, D. L. Mobley, and J. D. Chodera, "Capturing non-local through-bond effects in molecular mechanics force fields i: Fragmenting molecules for quantum chemical torsion scans [article v1.1]," *bioRxiv*, 2022.

[197] J. C. Shelley, A. Cholleti, L. L. Frye, J. R. Greenwood, M. R. Timlin, and M. Uchimaya, "Epik: A software program for pKaprediction and protonation state generation for drug-like molecules," *Journal of Computer-Aided Molecular Design*, vol. 21, pp. 681–691, Dec. 2007.

[198] D. Bajusz, A. Rácz, and K. Héberger, "Why is Tanimoto index an appropriate choice for fingerprint-based similarity calculations?," *Journal of Cheminformatics*, vol. 7, p. 20, May 2015.

[199] B. Thapa and K. Raghavachari, "Accurate pKa Evaluations for Complex Bio-Organic Molecules in Aqueous Media," *Journal of Chemical Theory and Computation*, vol. 15, pp. 6025–6035, Nov. 2019.

[200] M. Stahn, S. Ehlert, and S. Grimme, "Extended conductor-like polarizable continuum solvation model (CPCM-x) for semiempirical methods," *The Journal of Physical Chemistry A*, vol. 127, pp. 7036–7043, Aug. 2023.

[201] D. Anstine, R. Zubatyuk, and O. Isayev, "Aimnet2: A neural network potential to meet your neutral, charged, organic, and elemental-organic needs," *ChemRxiv*, 2023.

[202] X. Pan, F. Zhao, Y. Zhang, X. Wang, X. Xiao, J. Z. H. Zhang, and C. Ji, "MolTaut: A Tool for the Rapid Generation of Favorable Tautomer in Aqueous Solution," *Journal of Chemical Information and Modeling*, vol. 63, pp. 1833–1840, Apr. 2023.

[203] Z. Liu, T. Zubatiuk, A. Roitberg, and O. Isayev, "Auto3d: Automatic generation of the low-energy 3d structures with ani neural network potentials," *Journal of Chemical Information and Modeling*, vol. 62, p. 5373–5382, Sept. 2022.

[204] Y. Wang, I. Pulido, K. Takaba, B. Kaminow, J. Scheen, L. Wang, and J. D. Chodera, "EspalomaCharge: Machine learning-enabled ultra-fast partial charge assignment," Feb. 2023.

[205] P. Bleiziffer, K. Schaller, and S. Riniker, "Machine Learning of Partial Charges Derived from High-Quality Quantum-Mechanical Calculations," *Journal of Chemical Information and Modeling*, vol. 58, pp. 579–590, Mar. 2018.

[206] Q. Zhang, F. Zheng, T. Zhao, X. Qu, and J. Aires-de-Sousa, "Machine Learning Estimation of Atom Condensed Fukui Functions," *Molecular Informatics*, vol. 35, no. 2, pp. 62–69, 2016.

[207] S. Magedov, C. Koh, W. Malone, N. Lubbers, and B. Nebgen, "Bond order predictions using deep neural networks," *Journal of Applied Physics*, vol. 129, p. 064701, Feb. 2021.

[208] E. Caldeweyher, C. Bauer, and A. S. Tehrani, "An open-source framework for fast-yet-accurate calculation of quantum mechanical features," *Physical Chemistry Chemical Physics*, vol. 24, pp. 10599–10610, May 2022.

[209] D. C. Hiener, D. L. Folmsbee, L. A. Langkamp, and G. R. Hutchison, "Evaluating fast methods for static polarizabilities on extended conjugated oligomers," *Physical Chemistry Chemical Physics*, vol. 24, pp. 23173–23181, Oct. 2022.