IMPERIAL COLLEGE LONDON

MPHIL THESIS

High-throughput virtual screening of molecules for photon conversion

---

Author:
Shomik VERMA

Supervisors:
Prof. Aron WALSH
Prof. David SCANLON

A thesis submitted in fulfillment of the requirements for the degree of Masters of Philosophy

in the

Walsh Materials Design Group
Department of Materials

October 11, 2021

Statement of Originality

I, Shomik VERMA, declare that this thesis titled, "High-throughput virtual screening of molecules for photon conversion" and the work presented in it are my own. I confirm that:

- This work was done wholly or mainly while in candidature for a research degree at this University.
- No part of this thesis has previously been submitted for a degree or any other qualification at this University or any other institution.
- Where I have consulted the published work of others, this is always clearly attributed.
- Where I have quoted from the work of others, the source is always given. With the exception of such quotations, this thesis is entirely my own work.
- I have acknowledged all main sources of help.
- Where the thesis is based on work done by myself jointly with others, I have made clear exactly what was done by others and what I have contributed myself.

Copyright Declaration

Abstract

Photovoltaics (PV) have emerged as a prominent technology to generate electricity from sunlight. However, traditional single-junction PV cells such as silicon, thin film PV, and perovskites suffer from an inherent efficiency limit of 33.7%. This is primarily due to two loss mechanisms: sub-bandgap losses, where photons with energy below the bandgap of the PV cell cannot be utilized, and thermalization losses, where photons with excess energy above the bandgap lose their excess energy to heat. Photon conversion materials can help overcome the detailed-balance limit by converting wavelengths of light into energies the solar cell can efficiently absorb. The two common mechanisms for photon conversion are triplet-triplet annihilation (TTA) up-conversion and singlet fission (SF) down-conversion. Several molecules have been shown to exhibit TTA or SF, but there could be cheaper or less complex molecules previously overlooked that would be suitable. To identify such chromophores, high-throughput virtual screening (HTVS) of large databases is required.

Both TTA and SF involve the singlet and triplet excited states of molecules, so knowing these excited state energies is critical. The central issue to HTVS is that limited excited state databases exist, and computational techniques for calculating excited state energies are time-consuming. This thesis aims to solve this issue with various approaches. First, triplet excited state energies are predicted with a machine learning (ML) model trained on a dataset of TD-DFT energies generated with active learning (AL) to ensure the training set size is optimized. While directly predicting energies with ML is fast, there are issues with accuracy and training time. The second approach calibrates a high-throughput computational chemistry method called xTB-sTDA against TD-DFT with ML. This ensures both high accuracy and low computation time. Finally, the third approach applies xTB-ML to a large dataset, using AL to actively suggest candidate chromophores for photon conversion.

Acknowledgements

The work reported in this thesis was conducted between October 2020 and September 2021 in the Walsh Materials Design Group in the Department of Materials at Imperial College London, with funding from the Marshall Scholarship. Computations were conducted on Imperial's Research Computing Service facilities, specifically the CX1/2 clusters.[1] This project would not have been possible without the support of several people, some of whom are listed below.

Thanks to Miguel Rivera for his help getting me set up, teaching me the fundamentals of excited state chemistry, and guiding me through how to run calculations. I would be lost without your help!

Thanks to Kyle Swanson for creating Chemprop and making it so easy to use. You were so gracious in answering my many questions and helping troubleshoot my code.

Thanks to Jiali Li for his vast expertise in computation, specifically in active learning. I learned so much from collaborating with you.

Thanks to all the wonderful members of the Walsh Materials Design Group for great discussions and community over the past year. Thanks specifically to Kazuki for his many insights and suggestions, several of which became sections of my thesis. Thanks to Liam for reading over my thesis and providing comments. Thanks to Alex for pointing me towards the INDT dataset. Thanks to Jarvist for pointing me towards GB-GA.

Thanks to Jahnvi Jain for her patience, love, and laughs while I've been in the UK. I couldn't do it without you!

Thanks to my parents for their support, advice, and for housing me during the pandemic.

Finally, thanks to my advisors, Prof. Aron Walsh and Prof. David Scanlon, for their vision and invaluable guidance throughout the project. Thanks for taking a chance on hiring someone with no computational chemistry experience! Aron, I really appreciate all your efforts in making me feel welcome and included in the lab, even while we've been virtual. I'm in awe of your intellect, time-management skills, and how you're so nice to everyone! It's been an honor working with you.

And thanks to the reader for spending time reading my work! I hope you enjoy.

Contents

List of Figures

List of Tables

List of Abbreviations

AES	**A**nisotropic **E**lectro**S**tatic
AL	**A**ctive **L**earning
AO	**A**tomic **O**rbital
AXC	**A**nisotropic e**X**change **C**orrelation
CC	**C**oupled **C**luster
CG	**C**ontracted **G**aussian
CM	**C**oulomb **M**atrix
CP	**C**hem**P**rop
DFT	**D**ensity **F**unctional **T**heory
DTNN	**D**eep **T**ensor **N**eural **N**etwork
EA	**E**lectron **A**ffinity
ECFP	**E**xtended **C**onnectivity **F**inger**P**rint
FOM	**F**igure **O**f **M**erit
GBSA	**G**eneralized **B**orn **S**urface **A**rea
GCN	**G**raph **C**onvolutional **N**etwork
GFN	**G**eometries, vibrational **F**requencies, and **N**on-covalent interactions
GGA	**G**eneralized **G**radient **A**pproximation
GTO	**G**aussian **T**ype **O**rbital
HF	**H**artree-**F**ock
HOMO	**H**ighest **O**ccupied **M**olecular **O**rbital
HTVS	**H**igh-**T**hroughput **V**irtual **S**creening
IC	**I**nternal **C**onversion
INDT	**IND**olonaphthyridine **T**hiophene
IP	**I**onization **P**otential
ISC	**I**nter-**S**ystem **C**rossing
k-CV	**k**-fold **C**ross **V**alidation
KS	**K**ohn-**S**ham
LCAO	**L**inear **C**ombination of **A**tomic **O**rbital
LDA	**L**ocal **D**ensity **A**pproximation
LUMO	**L**owest **U**noccupied **M**olecular **O**rbital
MACCS	**M**olecular **ACC**ess **S**ystem
MAE	**M**ean **A**bsolute **E**rror
MCS	**M**aximum **C**ommon **S**ubstructure
ME	**M**ean **E**rror
ML	**M**achine **L**earning

MOPSSAM	**M**apping the **O**ptoelectronic **P**roperty **S**pace of **S**mall **A**romatic **M**olecules
MPNN	**M**essage **P**assing **N**eural **N**etwork
MSE	**M**ean **S**quared **E**rror
NIR	**N**ear-**I**nfra**R**ed
NN	**N**eural **N**etwork
PCQC	**P**ub**C**hem **Q**uantum **C**hemistry
PV	**P**hoto**V**oltaics
QBC	**Q**uery **B**y **C**ommittee
QM	**Q**uantum **M**echanics
RF	**R**andom **F**orest
RMSE	**R**oot **M**ean **S**quared **E**rror
S1	**S**inglet **1**st excited state energy
SD	**S**tandard **D**eviation
SF	**S**inglet **F**ission
SMILES	**S**implified **M**olecular **I**nput **L**ine **E**ntry **S**ystem
SOC	**S**pin-**O**rbit **C**oupling
sTDA	**S**implified **T**amm **D**ancoff **A**pproximation
T1	**T**riplet **1**st excited state energy
TD-DFT	**T**ime-**D**ependent **D**ensity **F**unctional **T**heory
TDA	**T**amm **D**ancoff **A**pproximation
TQY	**T**riplet **Q**uantum **Y**ield
TTA	**T**riplet-**T**riplet **A**nnihilation
TTET	**T**riplet-**T**riplet **E**nergy **T**ransfer
UCQY	**U**p-**C**onversion **Q**uantum **Y**ield
VR	**V**ibrational **R**elaxation
WFT	**W**ave**F**unction **T**heory
XC	e**X**change **C**orrelation
xTB	e**X**tended **T**ight **B**inding

Chapter 1

Introduction

1.1 Motivation

Solar energy technologies have garnered immense interest over the past several years,[2–5] in many countries around the world.[6–11] Specifically, extensive research has been conducted in the field of solar photovoltaics (PV), the process of converting sunlight directly into electricity.[12–17] A variety of PV technologies have been recently developed to go beyond conventional silicon solar cells, including perovskites,[18,19] thin film cells,[20] and organic cells.[21] However, all such solar cells suffer from an efficiency cap known as the detailed-balance limit, which states the maximum efficiency of a single-junction solar cell is 33.7%.[22]

The main loss mechanisms are known as spectrum losses, and can be categorized into sub-bandgap and thermalization losses. Sub-bandgap losses are inherent to solar cells as any photon with an energy below the bandgap cannot be absorbed by the cell. Similarly, thermalization losses inherently occur as any photon with excess energy above the bandgap loses this energy to heat upon absorption. Only considering spectrum losses, a single-junction solar cell would have a maximum efficiency of around 50%,[23] so reducing spectrum losses is crucial to increasing PV efficiency.

Several next-generation technologies have recently emerged to reduce such spectrum losses, such as multi-junction cells,[24] concentrated solar,[25] and hot carrier capture.[26] However, these processes often require expensive materials that may be difficult to manufacture.[27] Of interest to this work is the process of photon conversion, which would convert unusable or non-ideal wavelengths of light into energies the solar cell can absorb efficiently.[28] Photon conversion materials can be used with existing solar cell materials, limiting the need for novel infrastructure to be developed.[29] The following section outlines the basics of photon conversion and some common techniques to achieve conversion.

1.2 Overview of molecular photon conversion

Photon conversion technologies can be split into two groups: up- and down-conversion. As suggested by their names, up-conversion converts 2 low-energy photons

into 1 high-energy photon, while down-conversion does the opposite.

Several mechanisms exist for up-conversion, including photon avalanche,[30] excited-state absorption,[31] energy transfer up-conversion,[32] energy pooling,[33] and thermal upconversion.[34] However, these processes often require high-intensity, coherent light, such as concentrated solar power or even lasers, as well as typically expensive materials.[35] An up-conversion technique that works for low-intensity light in organic molecules is known as triplet-triplet annihilation,[36] which will be explored in this study.

Similarly, mechanisms for down-conversion include spontaneous parametric down-conversion,[37] quantum cutting,[38] and multiple exciton generation.[39] While some of these processes do indeed work at low intensities, they all require expensive inorganic materials such as lathanides or quantum dots. Down-conversion can occur in organic materials in a process called singlet fission,[40] of interest in this study.

Before discussing the details of singlet fission and triplet-triplet annihilation, it is important to understand some fundamentals of molecular luminescence. Figure 1.1 shows a Jablonski diagram demonstrating the basic phenomena of molecular luminescence.

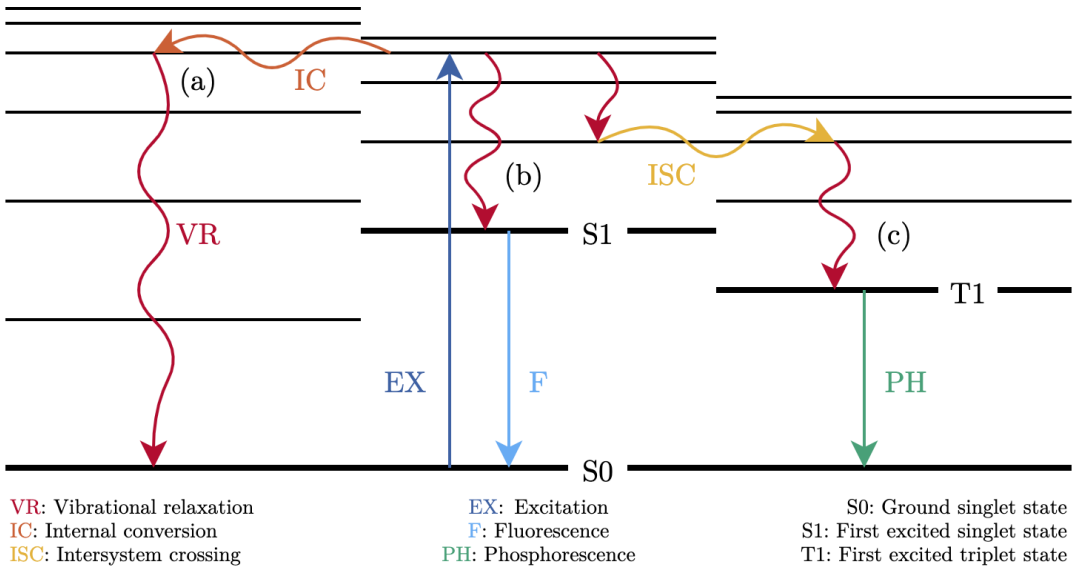


FIGURE 1.1: Jablonski diagram showing the basic phenomena in molecular luminescence. Processes grouped into (a) non-radiative decay, (b) fluorescence, and (c) phosphorescence. Straight arrows represent photonic processes while curved arrows represent internal electronic processes. Thick lines indicate energy states while thin lines indicate vibrational levels.

The luminescence process starts with excitation to an excited state energy level, as shown by the blue arrow, indicating excitation from S0 to S1. From here, there are typically three options. First, the system can undergo internal conversion (IC, orange arrow), from the S1 state into a high vibrational level of the S0 state, and non-radiatively decay (red arrow) into the ground state. This is shown as (a) in Figure

1.1, and no photon is emitted in this case. Second, the system can undergo vibrational relaxation (VR, red arrow) to the lowest vibrational level of the excited state, and then radiatively relax to the ground state (light blue arrow) through photon emission. This process is known as fluorescence, and is shown as (b) in Figure 1.1. Lastly, the system can undergo intersystem crossing (ISC, yellow arrow), in which the excited singlet state transforms into an excited triplet state, which can then radiatively relax (green arrow). This process is known as phosphorescence (shown as (c) in Figure 1.1), and typically a lower-energy photon than through fluorescence is emitted.

Note here some properties of the triplet excited state. As a triplet state, the electrons are unpaired such that they have parallel spin – typically, such an excited state cannot be populated directly from the ground singlet state because it is spin forbidden. However, in systems with strong spin-orbit coupling (SOC), ISC can occur, in which an electron in the excited singlet state reverses its spin. The triplet excited state is often favorable because of its long lifetime compared to the fast decay of an excited singlet state, allowing longer exciton diffusion lengths.[41]

1.2.1 Excited state energies

It is useful here to clarify some terminology for excited states. First, there is a fundamental difference between excitation and emission, although both involve energy state transitions within a molecule. Excitation involves a transition between the ground state and any vibrational level of an excited state. In contrast, emission involves a transition between the lowest vibrational level of an excited state and any vibrational level of a lower-energy state (e.g. the ground state). Figure 1.2(a) shows a Jablonski diagram of excitation vs. emission to help elucidate this concept.

From Figure 1.2(a), it may seem that any vibrational level in either energy state is easily accessible. In physical molecules, however, this is not the case, due to the Franck-Condon principle. This principle states that excitation and emission operate on much faster timescales than molecular geometry reconfiguration, so transitions are more likely to occur between states with equivalent configuration, even if such a configuration is not optimal for that energy state. This is shown in Figure 1.2(b), indicating that $0 \rightarrow 1$ excitation from S0 to S1 is more likely than $0 \rightarrow 0$. The same occurs for emission from S1 to S0. This creates a difference between peak excitation and peak emission, as shown in Figure 1.2(c), known as the Stokes shift. Measuring the S1 energy is therefore somewhat difficult, as the actual $0 \rightarrow 0$ energy transition (yellow lines) is unlikely, so care should be taken when lifting experimental values of S1 from literature to note how S1 was measured.

These fundamental processes of molecular luminescence form the basis of singlet fission and triplet-triplet annihilation technologies, as discussed in more detail in the following subsections.

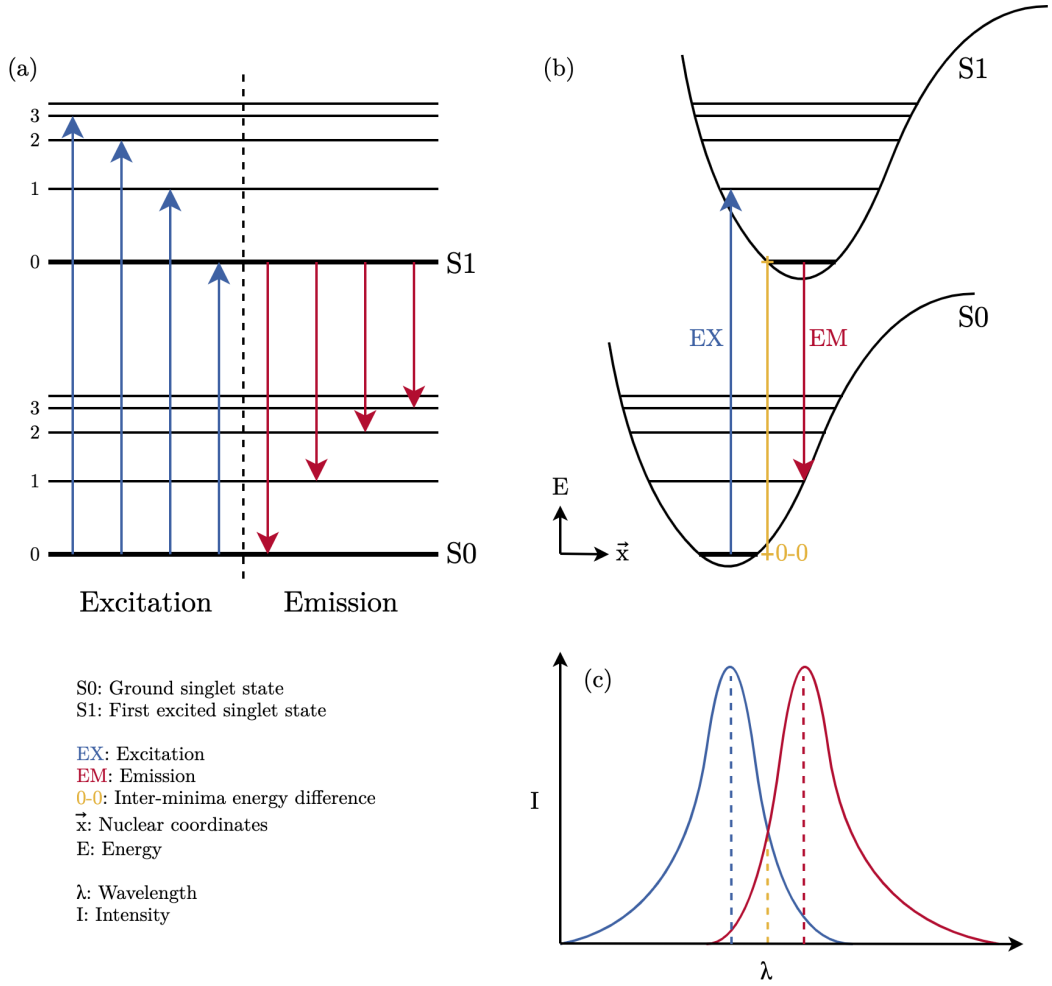


FIGURE 1.2: Schematics depicting the basics of excitation and emission. (a) Typical Jablonksi diagram of excitation and emission from ground state (S0) to excited state (S1), with various vibrational levels (0-3) depicted for both states. (b) Demonstration of Frank-Condon principle of 0→1 vertical excitation (blue arrow) followed by nuclear re-configuration and 1←0 vertical emission (red arrow). Also shows the 0→0 transition energy in yellow. (c) Shows the expected experimental excitation/absorption curve (blue) and emission curve (red), along with the theoretical 0→0 energy difference (dashed yellow line), demonstrating the Stokes shift.

1.3 Singlet fission

Singlet fission (SF) is the process of converting one high-energy photon into two low-energy photons. Specifically, a molecule excited into its singlet excited state transfers approximately half of its energy into a nearby molecule, exciting it into its triplet excited state, while simultaneously relaxing into its own triplet excited state. As mentioned before, the singlet to triplet transition is spin-forbidden, but SF circumvents this issue by coupling the generated triplet states as one singlet state. Thus, instead of relying on SOC and ISC, SF is instead classified as internal conversion (IC), and can compete with vibrational or radiative relaxation speeds.

1.3.1 Singlet fission materials

In 2010, a comprehensive overview of SF materials was presented by Smith and Michl.[40] As seen, conventionally, most SF materials have been molecular crystals[42] or aggregates,[43] as these allow rapid triplet diffusion and therefore more separation before phosphorescence. However, more recently, molecular dimers have emerged as potential SF materials as well.[44] These are interesting as solids require molecular packing so the versatility of structures is limited, while dimers are often in solution and can be more diverse. However, dimers usually exhibit low triplet quantum yields (TQY) (<10%), as seen in Table 2 of Smith and Michl.[40] Note that the maximum TQY is 200% as each singlet generates 2 triplets.

Fortunately, since that review was published, several papers have shown high TQY of oligomers and single molecules, as presented in Table 1.1

TABLE 1.1: Triplet quantum yield of recent singlet fission molecules and oligomers

Material	**Notes**	**Excitation Energy (eV)**	**Emission Energy (eV)**	**TQY (%)**
Perylene dimer[45]	Cofacial	5.00	2.05	56
Pentacene dimer[46]	3 isomeric configs	1.84	0.77	156
Pentacene dimer[47]	Polyaromatic encapsulation			196
Pentacene dimer[48]	Cross-conjugated			162
TIPS pentacene[49]	Single molecule			200
Tetracene tetramer[50]	Linearly linked	2.43	1.20	128
TIPS tetracene[51]	Single molecule			120

Clearly there has been significant improvement in solution-based singlet fission. However, most of these molecules are polyacenes, so there is still limited diversity of molecules. There has recently been work in designing new types of singlet fission materials,[43,52,53] which may reveal promising new directions.

1.4 Triplet-triplet annihilation

Triplet-triplet annihilation (TTA) combines two low-energy photons into one high-energy photon. While this process sounds similar to SF, it is significantly more involved. Eventually, two triplet states do combine to form a higher-energy singlet state (which is the opposite of SF), but the triplet states must first be generated. Thus, TTA requires a sensitizer molecule for triplet generation.

This occurs through ISC - the sensitizer's singlet excited state is first populated, which then transitions to a triplet state through ISC. Then, the sensitizer transfers its triplet state to the emitter, through a process called triplet-triplet energy transfer (TTET), a type of Dexter energy transfer (DET). DET is an energy transfer mechanism that requires no direct chemical bonding, rather it is dependent on wavefunction and spectral overlap. After the emitter is excited into its triplet state, when it encounters another triplet-excited emitter, it transfers its energy (like reverse SF) to generate an excited singlet state in the second emitter. This excited state can then fluoresce at a higher energy than the two input photons. Figure 1.3 provides a schematic for the TTA process.

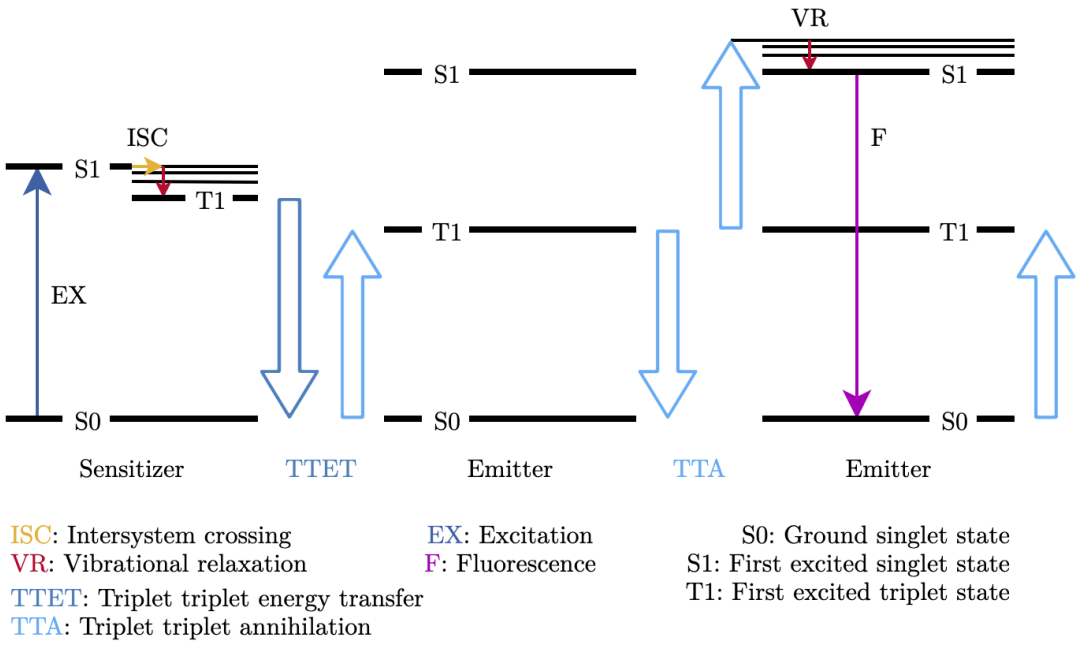


FIGURE 1.3: Overview of the TTA process. Sensitizer excitation generates S1 state, ISC transforms S1 into T1 state, TTET transfers sensitizer T1 to emitter T1, TTA between two excited emitter T1 states generates emitter S1 state, and emitter fluorescence emits a photon. Solid arrows indicate photon processes or internal electronic processes. Block arrows indicate inter-molecular energy transfer. Thick black lines are energy states while thin black lines are vibrational levels.

As seen, two low-energy photons are absorbed, and after several steps of energy transfer, eventually one high-energy photon is emitted. Among these steps are several loss mechanisms that reduce the energy of the re-emitted photon. First is the singlet-triplet split in the sensitizer. Because the triplet energy level is usually lower

than the singlet energy, there is a loss due to ISC and the resulting vibrational relaxation. Second, the triplet energy of the emitter may be less than the triplet energy of the sensitizer, creating a loss in the TTET process. Lastly, the singlet energy of the emitter may be lower than twice its triplet energy, creating further losses through vibrational relaxation. These three losses can be thought of as energy losses, reducing the final energy of the photon emitted in comparison to the sum of the energies of the absorbed photons. There are also efficiency losses, for example the oscillator strength of the singlet excitation, the ISC/TTET/TTA probability, and the photoluminescent quantum yield of the emitter, all of which dictate the probability of an absorbed photon being re-emitted. There are many current TTA materials, but due to the loss mechanisms above, their efficiencies are relatively low.

1.4.1 TTA materials

Simon and Weber,[36] Zhao et al.,[35] and Ye et al.[54] give comprehensive overviews of conventional TTA materials. See Table 1 in Ye et al.[54] and Figure 3 in Simon and Yeber[36] for some examples of typical sensitizer and emitter combinations. Aromatic molecules (such as perylene, 9,10-diphenylanthracene, rubrene, pyrene, etc.) are typically used as emitters, while metal-organic complexes (Pd and Pt porphyrins such as octaethylporphyrins and tetraphenyltetrabenzoporphyrins) are conventional sensitizers as they have strong SOC. Most conventional TTA materials have an up-conversion quantum yield (UCQY) between 10-20%.[55] To improve this efficiency, many novel materials are still being proposed and investigated.[56]

1.4.2 Near-IR TTA materials

In addition to efficiency improvements, there are also efforts to expand the wavelength regions TTA materials operate in. Currently, there are only certain classes of molecules that work well for near-IR to visible up-conversion. See Table 1 of the review published by Bharmoria et al. for a list of such materials.[57] Most materials still use porphyrins as sensitizers, tuned to specific absorption wavelengths, and aromatics as emitters. However, new materials such as quantum dot sensitizers, osmium complexes for direct triplet sensitization, or lanthanide-organic complexes are also being actively explored.[57] Note, however, that most of these materials have extremely low UCQY of <1%. Some of the TTA materials with the highest UCQY (>3%) are listed in Table 1.2, along with their excitation and emission energies in eV, adapted from Bharmoria et al.[57] Note all were measured in a toluene solution unless otherwise stated.

As seen, there are several limitations in current materials. The UCQY is low, the emission to excitation energy ratio is often significantly less than 2, and the excitation energy is not low enough to be useful for silicon solar cells (1.1 eV). Thus, there is clearly significant room for improvement in TTA performance, especially for near-IR to visible upconversion. While it is possible to synthesize new molecules and

TABLE 1.2: Upconversion quantum yield and energy deltas of highest-performing TTA materials.[57]

Material	Excitation Energy (eV)	Emission Energy (eV)	UCQY (%)
($PdPh_4OMe_8TNP$) / bis(phenyltetracene)[58]	1.78	2.49	4.0
($PdPh_4OMe_8TNP$) / BPEN[59]	1.78	2.18	3.2
PtTPTNP / PDI[60]	1.80	2.14	3.0
PtTPTNP / rubrene[60]	1.80	2.21	3.3
PbS-CdS / 5-CT(T) / rubrene[61]	1.53	2.21	4.2
$Os(tpy)_2^{2+}$ / $(i\text{-}Pr_2SiH)_2An$ in THF[62]	1.71	2.99	5.5
PdPc / rubrene[63]	1.70	2.21	5.6
PtPc / rubrene[63]	1.70	2.21	4.9
PbS- rubrene– DBP(E) as a crystal[64]	1.53	2.03	3.5

experimentally measure their efficiency, this is time- and resource- consuming. Additionally, synthesized molecules would likely follow the same classes of molecules that are known to perform well, rather than trying novel molecular structures. To expand the breadth of potential TTA (and SF) molecules, computational chemistry is needed.

1.5 Thesis aims

This thesis aims to use computational chemistry for high-throughput virtual screening (HTVS) of molecules for TTA and SF applications. There are several approaches taken to accomplish this, using a variety of methods as outlined in Chapter 2.

The easiest approach would be to screen already-existing databases with excited state properties calculated with high-accuracy computational chemistry techniques. Unfortunately, while excited state databases do exist, the ones including triplet energies are either small or only include certain types of molecules. Chapter 3 of this thesis will focus on generating a triplet dataset derived from PubChemQC (PCQC),[65] but instead of conducting millions of high-accuracy calculations, will focus on utilizing active machine learning to accurately and efficiently predict triplet energies.

While machine learning (ML) is a useful tool, there are a few drawbacks including potentially low accuracy and its inherent black-box nature which limits chemical

intuition development. An alternative to direct ML is calibration ML, the idea of calibrating high-throughput computational chemistry techniques against high-accuracy techniques, to achieve both fast computation and higher accuracy. As presented in the following section, the existing calibrations, while fast, do not increase accuracy significantly. However, there is precedent for using ML models to calibrate computational chemistry techniques against either higher-accuracy techniques or experiment. Chapter 4 will present a machine-learned calibration of xTB-sTDA against TD-DFT, termed xTB-ML.

Using xTB-ML, we can conduct HTVS of photon conversion molecules, with confidence in both accuracy and fast computation time. Chapter 5 discusses the application of active learning to sample the global chemical space of PCQC and suggest potential photon conversion molecules. Traditionally, active learning is paired with TD-DFT, but applying xTB-ML instead rapidly accelerates each active learning cycle.

Lastly, Chapter 6 will present some conclusions and avenues of future work.

The following Methods section will define some of the above terms and provide an overview of the methods used in this thesis.

Chapter 2

Methods

2.1 Computational chemistry for excited state calculations

As seen in Section 1.2, knowing the excited state energies of molecules is critical to designing new SF/TTA materials. There are various computational chemistry techniques used to calculate excited state energies (refer to Figure 1.2 for the various terms relevant to excited state energies). Excitation energy is approximated in computation by vertical excitation. It is "approximated" because the computational program will use the global energy minimum instead of the lowest vibrational level energy. Vertical excitation is easy to calculate as no excited state geometry relaxation is required. Emission energy is approximated with 3 steps: vertical excitation, excited state relaxation, and vertical emission. Again, the global energy minimum is used for these steps. Another common calculation is the adiabatic excitation energy, the energy difference between the energy minima of the excited and ground states, which can approximate the true $0 \rightarrow 0$ energy transition as shown by the yellow line in Figure 1.2(b). This circumvents the last step of the emission energy calculation, so only 2 calculations are required.

In this thesis, only vertical excitations are used. This is due to time constraints – in high-throughput screening of large databases, adding the computational expense of excited state relaxation would be prohibitively slow. Further, the Stokes shift for rigid molecules should be small – for example, Fang et al. showed the difference between adiabatic and vertical excitation energies for 96 systems ranging from inorganic homodiatomics to cyclic non-aromatic compounds was small, with a standard deviation of 0.1 eV.[66] Because most molecules considered in this thesis are small, rigid, and aromatic, the vertical and adiabatic excitation energies should be reasonably close.

The following sections will review some computational techniques for calculating the vertical excitation energy (from now on, referred to as the excitation energy or excited state energy).

2.1.1 High-accuracy techniques

There have been several techniques developed to calculate excited state energies of molecules. The techniques can broadly be classified into wavefunction theory (WFT)

or density functional theory (DFT). The difference essentially is that WFT seeks to calculate the wavefunction through approximate solutions to the Schrödinger equation, while DFT calculates the electron density which, from the Hohenburg-Kohn theorems, should uniquely determine the many-body wavefunction. Regardless, both are considered high-accuracy methods, and some specific techniques are outlined below.

The Hartree-Fock (HF) ab initio method is a basic WFT that other techniques have sought to improve upon. Such post-HF methods include configuration interaction (CI), electron propagator methods such as ADC(2), multiconfigurational self-consistent field (MCSCF) with its variant CASSCF, coupled cluster (CC) methods, or multireference (MR) methods such as MR-CISD or MRCC.[67] While these methods are very accurate and are often considered as reference data to evaluate accuracy of other techniques, they come with several challenges, perhaps most importantly a high computational cost, as well as particular protocols in identifying the active space that must be tuned for each molecule.[67]

Since WFT methods attempt to find the many-body wavefunction directly, they can be computationally expensive. Alternatively, for more computationally efficient calculations, DFT is often used. DFT finds many 1-electron wavefunctions and calculates the electron density from these individual wavefunctions. The electron density $n(\mathbf{r})$ can then be mapped to the total energy of the system E:

$$E[n] = T[n] + \int v_{ext}(\mathbf{r})n(\mathbf{r})d\mathbf{r} + E_H[n] + E_{XC}[n] \tag{2.1}$$

using the universal functional $F[n]$:

$$F[n] = T[n] + E_H[n] + E_{XC}[n] \tag{2.2}$$

where v_{ext} is the external (nuclear) potential, T is the kinetic energy, and the Hartree (Coulomb) energy E_H and exchange-correlation energy E_{XC} encompass the electron-electron interactions for the N interacting electron system. To make the problem more tractable, we can reformulate to a system of N non-interacting electrons, known as the Kohn-Sham (KS) equations:

$$\left(-\frac{\nabla^2}{2} + v_s(\mathbf{r})\right)\varphi_i(\mathbf{r}) = \varepsilon_i\varphi_i(\mathbf{r}) \tag{2.3}$$

where $\varphi_i(\mathbf{r})$ are the KS orbitals. v_s can be expanded as

$$v_s[n](\mathbf{r}) = \left[v_{\text{field}}(\mathbf{r}) + \sum_i^{N_Z}\frac{Z_i}{\mathbf{R}_i - \mathbf{r}}\right] + \left[\int \frac{n(\mathbf{r}')}{\mathbf{r} - \mathbf{r}'}d\mathbf{r}'\right] + \left[\frac{\delta E_{XC}[n]}{\delta n(\mathbf{r})}\right] \tag{2.4}$$

where N_Z is the number of nuclei with coordinates R_i and charge Z_i and E_{XC} is the exchange-correlation functional. This expansion includes the external (i.e.

nuclear) potentials, Hartree potential (electron-electron Coulomb potential) and exchange correlation. The exact form of E_{XC} is unknown, and must be approximated.

There are various levels of approximations employed for E_{XC}. With increasing order of complexity, they are the local density approximation (LDA), generalized gradient approximation (GGA), and hybrid functionals. Hybrid functionals combine some WFT-calculated exchange with LDA and GGA methods for a more accurate functional. The hybrid functional B3LYP[68] is the most commonly used functional in chemistry.[69] It contains 3 experimentally fitted parameters and combines LDA and the Becke 1988 functional (B88)[70] for exchange, and LDA and the Lee-Yang-Parr functional (LYP)[71] for correlation.[68]

For computational simplicity, the form of the KS orbitals are defined as a linear combination of basis functions $G_\alpha(\mathbf{r})$:

$$\varphi_i(\mathbf{r}) = \sum_{\alpha=1}^{N_{BF}} C_{\alpha i} G_\alpha(\mathbf{r}) \tag{2.5}$$

The simplest basis functions are atomic orbitals (AOs), but they can be generalized and optimized for computation by using Gaussian Type Orbitals (GTOs). A linear combination of GTOs $g_\nu(\mathbf{r})$ is called a contracted Gaussian (CG):

$$G_\alpha(r) = \sum_{\nu=1}^{N_\alpha} c_\nu g_\nu(\mathbf{r}) \tag{2.6}$$

Various levels of CGs are available as basis sets: STO-nG includes 1 CG composed of n GTOs per atomic orbital, 6-31G[72] includes 1 CG with 6 GTOs for core atomic orbitals and 2 CGs (1 with 3 GTOs and 1 with 1 GTO) for valence orbitals, 6-31G*[73] additionally includes polarization functions, and 6-31+G[73] includes diffuse functions. Other examples of basis sets are aug-cc-pVTZ[74] and def2-SVP.[75]

Once the 1-electron Kohn-Sham orbitals $\varphi_i(\mathbf{r})$ are known, the electron density $n(\boldsymbol{r})$ can be calculated as:

$$n(\mathbf{r}) = 2 \sum_{j}^{N/2} |\varphi_i(\mathbf{r})|^2 \tag{2.7}$$

and can then be back-propagated to find the total energy of the system.

To calculate excited state energies, the time-dependent version of DFT (TD-DFT) must be used.[76] Many of the formulations of DFT have analogues in TD-DFT: the Hohenburg-Kohn theorem is replaced by Runge-Gross, and the exchange-correlation functional is instead the exchange-correlation kernel. While the details of the TD-DFT implementation are beyond the scope of this section, the critical part is that linear-response TD-DFT allows calculation of excited states using the ground state density. Linear-response TD-DFT improves the zeroth-order approximation of ground-state transitions to generate true optical transitions. In addition to vertical transitions, TD-DFT also allows excited-state relaxation and calculation of adiabatic excitation energies.

In this study, the B3LYP functional is used. This is because a wide variety of molecules are considered, and B3LYP would be the most applicable to this diversity of structures. For DFT calculations, the 6-31G* basis set is used, while for the TD-DFT portion, 6-31+G* is used, to accommodate the potentially diffuse orbitals in the excited state. These (TD-)DFT settings are consistent with PubChemQC's workflow, which was another large-scale excited state study.[65] Larger basis sets and more complex functionals were avoided due to the high-throughput nature of this work.

Generally, TDDFT is considered a standard for excited-state calculations,[76] although its accuracy is not as high as WFT methods, especially for non-adiabatic dynamics such as bond breaking or conical intersections.[77] However, TD-DFT is often the go-to method for excited state calculations as it is significantly cheaper than WFT methods. TDDFT can be further simplified with the Tamm-Dancoff approximation (TDA), reducing computation time.[78]

However, despite being less expensive than post-HF ab initio methods, TDDFT (even with TDA) is relatively slow. Fortunately, several excited state databases already exist with calculations of molecular properties completed with either post-HF or TD-DFT methods. These will be described in the following subsection.

Excited state databases

There are several large databases of molecules, such as GBD-17[79] (166B) and PubChem[80] (100M). However, these are often missing crucial quantum chemistry data. The largest quantum chemistry dataset is QCArchive[81] (47M), which is a repository of various datasets such as ANI-1[82] (22M) and QM9[83] (134k). Unfortunately, excited state calculations are often not included in these datasets, as they require extensive additional TD-DFT calculations. There are, however, a few excited state databases, as outlined below.

QM7b[84] expands QM7[85] to include properties such as excitation energy calculated at the ZINDO,[86] SCS,[87] PBE0,[88] and GW[89] levels of theory for 7211 molecules. (QM7 itself is a subset of GDB-13[90] (970M) totalling 7165 molecules.)

QM8[91] is a subset of QM9 limited to 8 CONF atoms, totalling 21.8k molecules, with ground state, vertical excitation, and adiabatic excitation energies calculated using TD-DFT with the PBE0 functional and def2-SVP basis set. QM8 also includes calculations with the coupled-cluster RI-CC2[92] method using the def2-TZVP basis set.

Perhaps the largest excited state database is PubChemQC (PCQC),[65] containing the first 10 singlet vertical excitation energies for 3.5M molecules. The 3.5M molecules are a subset of all 100M PubChem molecules, without mixtures, isotopes, molecules with a period in their SMILES representation, molecules with elements Z>30, and charged molecules.[65] PCQC uses TD-DFT with B3LYP/6-31G* for ground state geometry optimization and B3LYP/6-31+G* for excited state calculations.[65] Note this format of (functional/basis set) will be used throughout this thesis to describe the level of theory used for TD-DFT calculations.

While the above databases only have singlet energies, there are also a few that include triplet energies, which as discussed above are relevant for photon conversion processes. VERDE materials DB (VerdeDB)[93] is a recent database consisting of 1.5k molecules relevant for renewable energy and green chemistry research, specifically including π-conjugated organic molecules such as porphyrins, quinones, and dibenzoperylenes. It includes singlet and triplet energies for 1k molecules.[93] VerdeDB uses TD-DFT with M06/6-31+G(d,p)[94] calculations for ground and excited state geometry optimization, therefore calculating 0-0 adiabatic excitation energies.[93]

QM-symex[95] is a database of the first 10 singlet and triplet vertical excitation energies of 173k (rotationally) symmetric molecules, expanding the 135k QM-sym[96] database with 38k additional generated molecules and calculating excited state properties for all molecules. This database uses TD-DFT with B3LYP/6-31G(2df,p) for ground state geometry optimization and B3LYP/6-31G for excited state calculations.[95]

QMspin[97] includes 13k singlet and triplet carbene structures. The ground state geometry optimization is done with TD-DFT, using B3LYP/def2-TZVP, but all excited state calculations are completed with post-HF methods.[97] MRCISD+Q-F12/cc-pVDZ-F12 is used to calculate the vertical spin gap while CASSCF(2e,2o)/cc-pVDZ-F12 is used for singlet and triplet excited state optimization, allowing 0-0 energy level calculations.[97]

Finally, QuestDB[98] contains high-quality calculations of singlet, triplet, and various other vertical excitation states on 500 molecules, using WFT methods for all calculations. Ground state optimization is done with CC3/aug-cc-pVTZ, and excited state calculations are done with the aug-cc-pVTZ basis set and a variety of WFT methods including CIS(D), ADC(2), CC2, and others.[98]

While these datasets are a useful resource, there are clearly limitations to the size and diversity of the constituent molecules, especially for triplet energies. To quickly calculate excitation energies of molecules not in one of these databases, it is necessary to turn to faster computational techniques, as outlined in the following section.

2.1.2 High-throughput techniques

Recently, work has been done in tight binding as an approximation to DFT to improve its computation time while retaining most of its accuracy. Specifically, density functional tight binding (DFTB)[99] was developed in the late 1990's[100] and featured a combination of the accuracy of DFT and the efficiency of semi-empirical quantum chemistry methods. However, the biggest drawback of DFTB is the extensive element pair-wise parameterization required, as well as the low transferability of the parameters.[99]

The eXtended Tight Binding (xTB) methods were designed to solve the issues with DFTB.[101] xTB methods generally feature optimized element-specific empirical parameters for $Z \leq 86$ and extended AO basis sets, while also including various energy terms such as classical repulsion, extended Hückel, and isotropic electrostatic

and exchange-correlation energy.[101] They differ from DFTB methods in that they utilize top-down parameterization, with semiempirical parameters fit to a large dataset rather than computed with first-principles calculations.[101] The first xTB method to be developed was GFN1-xTB,[102] which had all of the properties above. GFN2-xTB[103] was released a few years later to include a multipole electrostatic treatment and a more advanced dispersion model. These methods are called GFN as they are fast, robust, and accurate in calculating Geometries, vibrational Frequencies, and Noncovalent interactions.[102] Bannwarth et al. provide an excellent overview of the xTB family of methods in their recent paper.[101]

The primary approximation applied in tight-binding methods is considering molecular orbitals to be a linear combination of atomic orbitals (LCAOs). For the xTB family of methods, a partially polarized, minimal valence basis set with 1 CG composed of either 3 or 6 GTOs is used for each AO.

Further approximations are used when mapping the density to total energy. The total energy expression is similar to Equation 2.1, with LDA used for E_{XC}, and adding a non-local correlation (disperson) term. Instead of directly calculating the converged density $n(\mathbf{r})$, tight-binding methods use a reference density $n_0(\mathbf{r})$ composed of a summation of reference densities of each atom: $n_0 = \sum_A n_0^A$. The reference density is then related to the converged density with a density difference term Δn such that $n = n_0 + \Delta n$. The total energy can then be Taylor expanded around $\Delta n = 0$ as:

$$E[n] = E^{(0)}[n_0] + E^{(1)}[n_0, \delta n] + E^{(2)}[n_0, (\delta n)^2] + E^{(3)}[n_0, (\delta n)^3] + \dots \tag{2.8}$$

Most tight-binding methods, including GFN(1,2)-xTB, truncate the expansion after 3 terms. After expanding and calculating these terms, we see specific physical processes expressed at each order. The zeroth order includes repulsion (E_{rep}) and dispersion (E_{disp}), first order includes an extended Hückel-type term (E_{EHT}) and dispersion again, second order includes electrostatic (E_{ES}), exchange-correlation (E_{XC}), and further dispersion, and third order includes E_{XC} and dispersion. Each GFNn-xTB method includes different terms for each order, for example, GFN2-xTB uses:

$$E = E_{rep}^{(0)} + E_{EHT}^{(1)} + E_{\gamma}^{(2)} + E_{AES}^{(2)} + E_{AXC}^{(2)} + E_{disp,D4}^{(2)} + E_{\Gamma}^{(3)} \tag{2.9}$$

where AES and AXC are anisotropic terms, E_Γ is an onsite electrostatic/exchange-correlation correction, and D4 signifies a modified D4 dispersion model. More details about the form of each energy term is available in Bannwarth et al.[101] These approximations (LCAO basis set and the third-order Taylor-expansion) allow fast yet accurate computation of ground state properties.

To conduct excited-state calculations, Grimme introduced the simplified Tamm-Dancoff density functional approach (sTDA)[104] as an approximation to TD-DFT. The theory behind sTDA, the specifics of the simplifications employed, and details of

the parameterization technique are available in the original paper.[104] The key approximations of sTDA include simplifications to two-electron integrals and setting an upper limit to the excitation space, which improve computation time by 2 orders of magnitude.[104] Note that because sTDA was developed to calculate excitation spectra, there is no excited state relaxation component, so only vertical excitation energies can be calculated.

In this study, we used GFN2-xTB for ground-state optimization with sTDA for excited state calculations, following the workflow presented by Grimme and Bannwarth in 2016, called xTB-sTDA.[105] We specifically used the `tight` threshold for geometry optimization, with the GBSA solvation model using benzene to mimic a non-polar environment. We then used the `xtb4stda` package to prepare the wavefunctions output by xTB for sTDA. sTDA then calculated excited-state properties, using an energy threshold of 10 eV.

Of relevance to this thesis are the computational time improvements provided by xTB-sTDA compared to TDDFT.[105] Specifically, excited state properties of several molecules with 100s of atoms were able to be computed in minutes, and molecules with 10s of atoms completed within seconds.[105] A more comprehensive analysis of computation time for xTB-sTDA compared to TDDFT was done in this work, as shown in Figure 2.1.

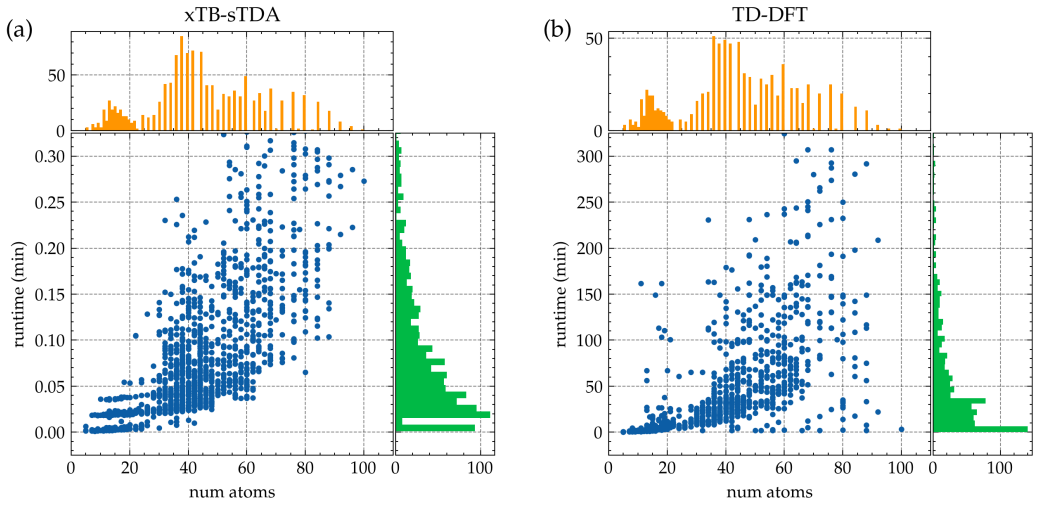


FIGURE 2.1: (a) xTB-sTDA vs. (b) TDDFT runtime comparison for S1 calculations of molecules in VerdeDB.[93] Center plots (blue datapoints) are scatter plots of computational runtime versus number of atoms in each molecule, while side plots (orange and green bars) show histograms to demonstrate the distribution of datapoints.

As seen, xTB-sTDA has immensely lower runtime compared to TD-DFT for excited state calculations, with potentially 3-4 orders of magnitude reduction. Most of this runtime reduction is due to the time savings of ground state optimization, as xTB takes 30-60 seconds, while DFT takes 3-4 hours. The time savings with using sTDA are less drastic, with sTDA also taking 30-60 seconds, while TD-DFT takes 5-10 minutes.

A natural tradeoff with faster computation time is potentially lower accuracy. In Grimme and Bannwarth's original paper introducing xTB-sTDA, they reported a mean absolute error (MAE) between sTDA-xTB and reference energies (calculated from SCS-CC2/TD-DFT) of between 0.34 and 0.48 eV, depending on the complexity of the input structure.[105] A further, more comprehensive analysis of sTDA error is presented in Chapter 4 of this thesis. Regardless, for many applications, especially for a first-pass screening for large databases, this level of error is acceptable. Due to the fast computation time and relatively low error of xTB-sTDA, many recent works have employed this methodology.

Recent works using xTB-sTDA

The publication introducing sTDA has been cited over 130 times, with studies applying sTDA to diverse systems including cyanobacteriochromes,[106] conjugated polymers,[107] porphyrinoids,[108] and proteins.[109] Many of these studies[106,108] attempt to benchmark sTDA. For example, Batra et al. compared combinations of different TD-DFT approximations (including sTDA), basis sets, and functionals against experimental reference values of excited state energies of porphyrinoids.[108] They found sTDA with the def2-SVP basis set and the CAM-B3LYP[110] functional to be the most accurate, with an MAE of 0.05 eV across 12 molecules.[108] Wiebeler and Schapiro compared various computational chemistry techniques, including sTDA, for structure and excited state prediction of the cyanobacteriachrome.[106] They found sTD-DFT and RI-ADC(2) using the CAM-B3LYP functional for the ground state optimization predicted experimental results well, but sTDA blue-shifted results by 0.14 eV.[106]

Note, however, that these studies only used sTDA, with a different computational chemistry technique (such as DFT or post-HF methods) to calculate the ground state structure. While this helped improve the accuracy of the method, the overall speed of calculation is reduced. In contrast, several other studies, mostly from Zwijnenburg and coworkers, used xTB with sTDA to achieve faster computation time for high-throughput screening, and they were able to successfully screen large databases of copolymers,[111] conjugated polymers,[107] small aromatic molecules,[112] photocatalysts,[113] and organic dyes.[114]

Using xTB instead of higher-accuracy methods for ground state structures naturally leads to greater errors. To retain relatively high accuracy while keeping the computation time low, instead of taking the raw xTB-sTDA excited state values, Zwijnenburg and coworkers calibrate the data against a few TD-DFT calculated values. For example, Wilbraham et al. desired to map the excitation energies of small aromatic molecules in chemical space with xTB-sTDA.[112] 143 molecules were used as a calibration set for a linear shift correction, and the linear shift improved the MAE of the 143 set from 0.25 to 0.21 eV.[112] The linear shift was then applied to all 250k molecules considered in the study.[112] Similarly, Heath-Apostolopoulos et al. explored the property space of diketopyrrolopyrrole dyes with xTB-sTDA.[114] They used a sub-set of 105 dyes for the linear calibration, but found a poor correlation

between xTB-sTDA and TD-DFT for sulfur-containining dyes.[114] Linear calibration was able to improve the error for sulfur-containing dyes from 0.20 to 0.18 eV.[114] While linear calibration is a quick method for calibration that requires minimal computational expense, there is clearly room for improvement.

Using methods such as machine learning to calibrate high-throughput methods against high-accuracy methods could help improve results. Machine learning could also be useful as a standalone computational chemistry technique trained on high-accuracy methods to directly predict desired properties. The following section outlines machine learning applied to excited state chemistry.

2.2 Machine learning for excited state chemistry

A thorough review of ML for molecular excited states is presented by Westermayr and Marquetand.[67] Most of the details in the review are beyond the scope of this thesis, besides a few key points.

First, relevant ML models to this thesis will directly predict the desired property (excitation energy or error between excitation energies). Westermayr and Marquetand refer to this as the tertiary output, considering the wavefunction/density as the primary and the energies as secondary output.[67] For high-throughput screening, however, having the ML model directly output the desired property is most efficient.

Next, we must consider which ML settings to use, including what type of model, how to transform molecular structure into a machine-learnable format, and what model architecture to use. Further, it would be useful to know how these different settings historically have performed in predicting molecular properties. The following subsections will detail these two points.

2.2.1 Common molecular machine learning implementations

At the most basic level, ML models for predicting excited states are most likely regression models. Namely, the ML model would be trained on a dataset of molecular structure information (SMILES, 3D geometry, etc) labeled with numerical values, and the ML prediction would output a number for a given molecule. This is in contrast with classification models that would be trained on a set of molecules labeled with categories, and then be used to predict the category of a test molecule.

The next step to generating an ML model is generating machine-learnable descriptors for each molecule, since models typically require a numerical representation of a molecule as input.[115] This can be done by featurizing the molecular representation at either the molecule, substructure, or atom level. Featurizers can range in complexity, including element fractions,[116] distance or Coulomb matrices,[117] substructure fingerprints,[118] or more complex graph and matrix features.[118]

Once the molecule has been featurized, it can be used as input for an ML model. The most common architecture used for molecular ML is the neural network (NN).

NNs work by passing information through layers, inferring patterns and learning efficient representations of molecules with each successive layer.[119] NNs are widely used as they are highly flexible, have few hyperparameters to optimize, and have multi-task capabilities.[120]

There are various implementations of NNs, such as convolutional NNs (including graph (GCN) or text), message passing NNs (MPNNs), and deep tensor NNs (DTNNs).[118] GCNs have been used widely for molecular ML as molecules can be naturally represented as graphs.[121] MPNNs allow easy featurization of molecules as descriptors are passed as input parameters in the network architecture..[122]

NNs must be trained with data in order to make accurate predictions. The neurons in each layer of the NN have different weights for inputs, and the purpose of training is to optimize these weights. The training process occurs over several "epochs," as an optimization algorithm attempts to find the best weights to predict outputs given inputs. The metric used to determine performance during training is called "loss," and many loss functions exist including mean squared error (MSE):

$$MSE = \frac{1}{n}\sum_{1}^{n}(y_i - \hat{y}_i)^2 \tag{2.10}$$

where y_i is the ith prediction, $\hat{y}_i$ is the ith true value, and n is the number of datapoints. Root mean squared error (RMSE) is another common metric, and is simply the square root of MSE.

To increase accuracy of results, it is possible to create an ensemble of ML models and average predictions. There are several methods of creating ensembles. The simplest is to train multiple models on the same dataset – due to the random initial weights defined and the often stochastic nature of the weight optimization algorithm, the model will output slightly different results each time. Another technique is data resampling, or choosing subsets of a larger dataset to train multiple ML models. This can be done either randomly, or with k-fold cross-validation (k-CV). k-CV splits the data evenly into k folds, and k times, one fold is set aside as non-training data while the model is trained on the remaining data. Thus, k ML models are generated, each with slightly different training data. For all of the above techniques, the ensemble of models can be used to predict properties, and the average of predictions is often more accurate than any individual prediction.

In this study, a few NNs are considered, but the Chemprop MPNN[123] (CP-MPNN) is ultimately used. CP-MPNN operates on a 2D graph representation of a molecule with atoms as nodes and bonds as edges. CP-MPNN is novel as it focuses on bond-centered data rather than atom-centered, which limits extraneous loops which can create noise in the final representation.[123] The MPNN generates a neural representation of the molecule through multiple steps of directed message passing.[123] Each bond direction has a hidden state and a message, which are updated through T message passing update steps.[123] After the bonds are fully updated, the model transitions back to an atom representation by summing all incoming directed bonds

for each atom.[123] Then, adding the data for all atoms results in the final molecular featurization.[123] Finally, this featurization is fed to a feed-forward NN for property prediction.[123] CP-MPNN was compared to several ML models with various featurization techniques, and outperformed their performance for a wide variety of datasets and desired properties.[123]

In this study, CP-MPNN is used both for directly predicting excited state energies, and to calibrate high-throughput computational chemistry techniques to achieve higher accuracy. For ensembling, both random initial weights and k-CV are used.

While CP-MPNN has not yet been used to predict excited-state energies, other molecular ML models have predicted such properties with great accuracy over the last few years. The next subsection will give an overview of the performance of recent ML models for excited state energies.

2.2.2 Predictive performance of recent models

Relevant ML models predict excited state energies of molecules in large-scale databases. The following works fall under this category.

Montavon et al. predicted various ground- and excited-state properties (including S1) by training a deep, multi-task NN with a Coulomb matrix (CM) variant descriptor on 5k molecules randomly sampled from QM7b.[84] They then tested the ML model on the 2.1k remaining molecules, achieving an MAE for S1 of 0.13 eV, and an MAE for S_{max} of 1.06 eV.[84] The ML model required 3D structure as input, generated from OpenBabel[124] and PBE[125] for ground state optimization.[84]

Pronobis et al. tried to directly predict various TDDFT energies by training a kernel ridge regression model in combination with 2-body and 3-body interaction descriptors on 10k molecules sampled from QM8.[126] They then tested the ML model on the remaining unseen 11.8k molecules, achieving an MAE for S1 of 0.48 eV.[126] They also compared the ML model to a similarly trained SchNet DTNN, which achieved an MAE of 0.49 eV.[126] Similarly to the previous work, this study also used relaxed geometries (from QM8) as input.[126]

Ghosh et al. trained a DTNN to predict molecular excitation spectra.[127] They took the molecular coordinates and atomic charges of 132k molecules from QM9 as input, using the 16 highest PBE+vdW eigenvalues as excitation energies.[127] They tested their model on the 10k training set from Ramakrishnan et al.,[91] finding an average RMSE of 0.19 eV and an RMSE of 0.16 eV for the lowest excitation energy.[127]

Nakata and Shimazaki, in their paper introducing PCQC, also generated an support vector machine regression model with a radial basis function kernel to predict highest occupied molecular orbital (HOMO) - lowest unoccupied molecular orbital (LUMO) gap in molecules.[65] They simply used the SMILES representation of the molecule as input, featurized with a 1024-bit topological fingerprint.[65] They then trained the model on 20k randomly selected molecules and tested on 980k molecules, finding an RMSE of 0.36 eV.[65]

Finally, Kang et al. trained a random forest (RF) ML model to predict the excitation energy with the maximum oscillator strength for molecules in PCQC.[128] They again only used the SMILES strings as input, featurized with the extended-connectivity fingerprint (ECFP), Molecular ACCess System (MACCS) keys, and RDKit descriptors.[128] They then trained the RF model on 450k randomly sampled molecules, and tested the model on a randomly sampled 50k molecule test set, finding an an RMSE of 0.43 eV.[128]

All of the above models attempted to directly predict molecular properties using ML. Taking a different approach, Ramakrishnan et al., in their paper presenting QM8, used machine learning to correct TDDFT values against CC2, instead of directly predicting excited state values.[91] They trained a kernel model using molecular geometry as input with CM and bag-of-bonds (BOB) descriptors on 10k molecules randomly sampled from QM8.[91] The delta ML model improved the MAE from 0.27 eV (using just TD-DFT) to 0.1 eV, when predicting S1 on the 11.8k remaining molecules.[91]

This type of ML model has many names (delta, deviation, calibration, correction, etc.), and has been used extensively in the past. Ramakrishnan and coworkers published another study in 2015 using the delta ML approach to predict "enthalpies, free energies, entropies, and electron correlation energies" of various molecules.[129] They attempted to correct the semi-empirical ZINDO method against the GW method, and tested their approach on 7k small organic molecules, using a training subset of 1k molecules.[129] They found the MAE decreased from 0.78 to 0.23 eV for HOMO and 0.91 to 0.16 eV for LUMO when using the ML model.[129] They then expanded their scope to predict enthalpies of 134k molecules, correcting PM7 baseline values against B3LYP target values, finding the MAE decreased from 7.2 to 3.0 kcal/mol with a 10k training set tested on the remaining 124k molecules.[129]

Recently, Pollice et al. developed a workflow for HTVS of organic molecules with inverted singlet-triplet splits, i.e. triplet excited state energy greater than singlet.[130] Such materials require high-accuracy methods at the post-HF ab initio level (such as coupled-cluster), but these are inherently incompatible with high-throughput workflows as they are too computationally expensive.[130] Pollice et al. got around this issue by calibrating ωB2PLYP/def2-SVP TD-DFT against EOM/CCSD using Gaussian process regression, finding a 200-fold time reduction with high accuracy after calibration.[130] They also calibrated computed vertical S1 excitation energies against experimental values from UV/Vis absorption data, using a linear regression, and found an R^2 around 0.9 for the shift.[130] Unfortunately, because calibration was not the primary objective of this work, they do not provide many details for either of these corrections, though they acknowledge that this is an avenue of future work.[130]

The group of G.H. Chen has done extensive work in calibration ML models to increase the accuracy of TD-DFT compared to experiment.[131–135] Hu et al. first

introduced the idea by calibrating heat of formation values calculated from TD-DFT against experiment for 180 small/medium organic molecules using a neural-network ML model.[131] They found an improvement in RMSE from 21.4 to 3.1 kcal/-mol with an ML model trained on 150 molecules and tested on 30.[131] Sun et al. improved this model in 2014 by adding sampling and bootstrapping methods and expanding the dataset size to 539, resulting in an improvement in MAE from 14.95 to 1.31 kcal/mol for a 90-molecule test set.[132] Yang et al. further improved the model for large molecules in 2018 with a new size-independent ML correction, with a MAE improvement from 28.75 to 1.67 kcal/mol for a test set of 13 molecules.[133]

Perhaps more relevant to this study, Wang et al. used a correction ML model to predict absorption energies, calibrating B3LYP/6-31G(d) data against experiment for 60 molecules. Training the model on 50 molecules and testing on 10, they found the RMSE reduced from 0.33 to 0.09 eV after using a neural-network based ML model.[134] Li et al. improved upon this model by adding a genetic algorithm component and expanding to 150 molecules (120 train/30 test). They again predicted absorption energies and calibrated TD-DFT against experimental values, reducing the RMSE from 0.47 to 0.16 eV.[135]

Unfortunately, most of these calibration models use fairly small datasets, as experimental data is often scarce. Regardless, as seen, there is significant precedent for using both direct ML models and delta ML models for calculating excited state data. Many of the above examples used randomly sampled training sets for the ML model, but it is possible to more purposefully generate a training set, through active learning, as discussed in the next section.

2.2.3 Active learning

All of the studies presented in the previous section use randomly generated training sets to train an ML model to predict excited state energies. However, this sampling technique may be inefficient, with oversampling of easily predicted regions of chemical space. Ff the reference data is generated with high-accuracy computational chemistry techniques, inefficiency can vastly increase computational time. At the same time, random sampling may also create inaccurate ML models, if certain important regions of chemical space are undersampled.

Active learning (AL) can help solve this issue. AL is the process of building up a training set piece by piece. Starting with an initial training set, an ML model is generated and used to determine which areas of chemical space require more sampling. This is done with a measurement of uncertainty, usually by creating an ensemble of individual ML models and taking the variance in predictions as uncertainty. Then, the highest uncertainty molecules can be labeled and added to the training set, and the process repeated.

Another application of AL is to suggest suitable molecules on the fly. Instead of using uncertainty as a metric to choose molecules, a suitability function is defined which quantifies how well a molecule matches desired properties. Then the ML

model can be run on the database and be used to find suitable molecules, which can be added to the training set to refine the model, and so on in cycles.

The concept of AL is derived from Bayesian optimization, which is a global optimization technique for functions that would be too expensive to evaluate directly. Instead, a surrogate function is created, uncertainty in the surrogate is evaluated, and an acquisition function based on the uncertainty and potentially other factors is used to guide sampling. Bayesian optimization was applied to ML in 1992 and termed "query by committee" (QBC).[136] In QBC-based AL, the surrogate is an ML model, uncertainty is usually ensemble variance, and the acquisition function is usually the sum of uncertainty and potentially suitability.

In this work, active learning was used both for training set generation and on-the-fly suggestion of candidate molecules. For training set generation, the acquisition function was solely based on uncertainty, and a training set was built up of high-uncertainty molecules. This was to generate the smallest dataset required to achieve low-error predictions. For molecule suggestion, the acquisition function included both uncertainty and suitability, allowing quick suggestions of candidate molecules. There is significant precedence for AL being used for molecular ML. The following studies are most relevant to this thesis.

Gubaev et al. used active machine learning to compose a dataset used to predict enthalpies of molecules in QM9.[137] Starting with a dataset of 1k randomly sampled molecules, they ran 22 cycles of AL to achieve a 6k set.[137] They developed a custom-defined "novelty" acquisition function to prevent similar molecules from being added to the training set, as well as a custom moment tensor ML model.[137] While the MAE and RMSE of the final AL training set was only slightly less than a randomly sampled 6k training set, the maximum error was significantly lower (20 vs. 160 kcal/mol), indicating AL did a much better job of limiting outliers.[137]

Kunkel et al. used active learning for the discovery of novel organic semiconductors in an unlimited search space with a Gaussian process regression ML model.[138] They started with an initial training set of 179 molecules, and defined an acquisition function that included both fitness and uncertainty, with a weighting parameter included to prioritize one over the other.[138] They first tested their AL on a fixed space of 65,552 molecules, out of which 2438 molecules were high-performing, finding that after 50 AL cycles they were able to consistently identify 70-80% of the high-performing molecules, with a training size of only 5179 molecules.[138] Applying AL to the unlimited chemical space, after 15 AL cycles, they generated a total of 1680 molecules, of which 900 had favorable characteristics.[138]

Smith et al. used active learning to improve prediction of molecular potential energy.[139] The previously generated ANI-1 model was based on random sampling, but in this work they generate ANI-1x using AL.[139] They first reduce ANI-1 to eliminate redundant molecules, then expand the training set by using AL to sample small molecules from GDB11, ChEMBL, and algorithmically generated dipeptides.[139] ANI-1x was able to match the performance of ANI-1 with 10% of its training size and

vastly outperform (by 5x) ANI-1 with 25% of its training size.[139]

Gómez-Bombarelli et al. reported a comprehensive virtual screening approach for organic LEDs, starting with a 1.6M molecular library and ending with experimental validation of candidate molecules.[140] Active machine learning was used as TD-DFT calculations on 1.6M molecules would be prohibitively slow.[140] The AL parameters were as follows: the initial training set was 40k randomly selected molecules, the surrogate model was a multi-task NN, and the acquisition function was a figure of merit based on the singlet-triplet split and oscillator strength, not considering uncertainty.[140] The model was used to suggest molecules for further analysis with TD-DFT, eventually suggesting 400k molecules, of which 3,000 had both high oscillator strength and low singlet-triplet split.[140] This is one of the few studies applying AL to excited state calculations, but because the focus was on materials discovery, unfortunately not many details are provided for AL implementation.

Clearly, active learning is a useful addition to conventional ML, as it can: reduce the size of the training set, improve the predictive power of the ML model generated, and suggest potential candidate molecules on the fly.

As seen, there are a variety of techniques that can be used for high-throughput virtual screening of molecules, including TD-DFT, xTB-sTDA, ML, and AL. The following chapters will apply these techniques in various forms to identify novel chromophores. Specifically, Chapter 3 will use TD-DFT, ML, and AL to directly predict excited state energies. Chapter 4 will use ML to calibrate xTB-sTDA against TD-DFT. Chapter 5 will use ML-calibrated xTB-sTDA paired with AL to rapidly screen molecules and identify candidates for TTA/SF. Each of these chapters will start by recounting some motivation, then outline the specifics of the methodology (including how the above methods were used in the workflow), present and discuss results, and finally summarize the work and outline some future directions. The final chapter (Chapter 6) will present the overall conclusions of this thesis.

Chapter 3

Active machine learning for triplet dataset generation

3.1 Motivation

As discussed in Section 1.2, knowing triplet excited states of molecules is critical for discovery of new materials for low-intensity photon conversion techniques such as singlet fission (SF) and triplet-triplet annihilation (TTA). While it is possible to experimentally measure triplet energies, this is time- and resource- consuming. It is much cheaper to conduct virtual screening of molecules with calculated triplet energies. However, very few triplet energy databases exist: of the databases outlined in Section 2.1.1, only 4 (VerdeDB,[93] QM-symex,[95] QMspin,[97] and QuestDB[98]) have triplet energies. Additionally, these databases have some restrictions: VerdeDB is small with only 1k triplet energies calculated,[93] QM-symex only includes symmetric molecules,[95] QMspin only contains carbene structures,[97] and QuestDB is also small with 500 molecules.[98] A larger database would offer depth and breadth: broad exploration of the chemical space for previously unknown classes of molecules that could serve as efficient photon converters, and deep exploration of a subset of chemical space for molecules with more suitable properties.

Ideally, we would expand PubChemQC (PCQC)[65] to triplet excited state data, as it already includes TD-DFT calculations for the first 10 singlet excited states of 3.5M molecules. However, generating TD-DFT triplet energy data for millions of molecules is computationally expensive: the PCQC project was started in December 2013 and only completed calculations for 2 million molecules in June 2015.[65] For triplet calculations, at least the optimized ground state structure has already been provided in PCQC, but still 3.5M triplet state calculations would take around 41 months to complete on the Imperial cluster, assuming each molecule takes 5 minutes to calculate and 10 jobs can run concurrently. Therefore, this study aims to use ML to generate triplet data for the 3.5M molecules in PCQC. Essentially, an ML model will be trained on a smaller training set (<10% of the dataset size) and used to predict triplet energies of the remaining molecules with reasonable accuracy.

Conventionally, a training set would be composed of randomly sampled molecules. The following section discusses this approach.

3.1.1 Conventional ML

For the conventional ML approach, 500k molecules were randomly sampled from PCQC and formed the training set. Only singlet energies were used, since triplet energies were unavailable in the database. The ML model architecture was a message-passing neural network (MPNN) generated using ChemProp (CP).[123] A CP MPNN ensemble model was trained with 10-fold cross-validation splits - meaning the data was split into 80%/10%/10% train, validation, test sets 10 times with unique test sets, and 10 ML models were trained based on the 10 subsampled training sets. Then, the ensemble model was used to predict the S1 energies for 350k (unique) molecules randomly chosen from PCQC. Figure 3.1 shows plots of the ML predictions versus the ground truth values taken from PCQC.

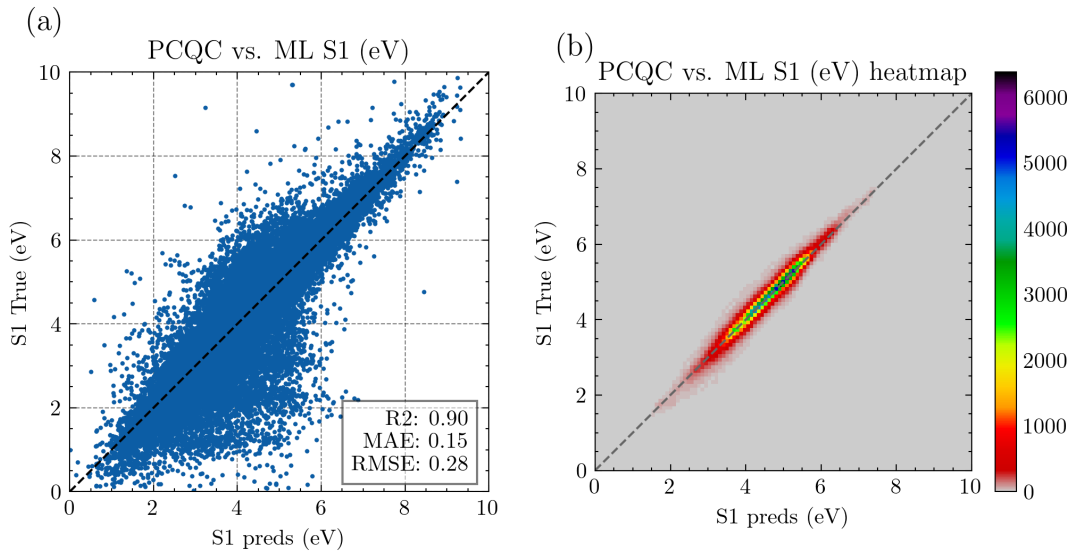


FIGURE 3.1: Comparison of ML-predicted S1 energies (x-axis) vs. TD-DFT generated S1 energies from PCQC (y-axis). (a) Shows all datapoints while (b) shows a heatmap of values, with the colorbar representing the number of molecules in each pixel. Inlaid box shows quantitative measurements of accuracy for ML predictions.

Here, it is evident from the R2 of 0.9 that the ML model accurately predicts the TD-DFT ground truth data. While the distribution of data is wide, the heatmap removes outliers and shows most of the data is in excellent agreement with TD-DFT values. A histogram of calculated errors is provided in Appendix Figure A.1. The MAE of 0.15 eV and RMSE of 0.28 eV are similar to those of previous models for excited state energy predictions, as presented in Section 2.2.2. Further, since a 500k-molecule trained ML model accurately predicts S1 energies, it is reasonable to assume a similarly sized dataset would accurately predict T1 energies.

However, 500k TD-DFT calculations would be slow even given an optimized ground state structure (approximately 6 months given each molecule takes 5 minutes to calculate, and the Imperial cluster allows 10 concurrent jobs). For this reason, it is necessary to generate a training set more intelligently using active learning (AL).

Section 3.2 discusses the AL methodology, presenting the workflow used in this work as well as optimizations to the workflow. Section 3.3 then presents the results, first testing the optimized AL workflow on singlet energies (since these are known), and once proven to work, applying the workflow to generate triplet data. Combining these two predictive models, Section 3.3.4 identifies candidate molecules for photon conversion.

3.2 Methodology

3.2.1 Active learning workflow

Active learning is introduced in Section 2.2.3 as a way to build up training sets. Figure 3.2 shows the AL workflow used in this work.

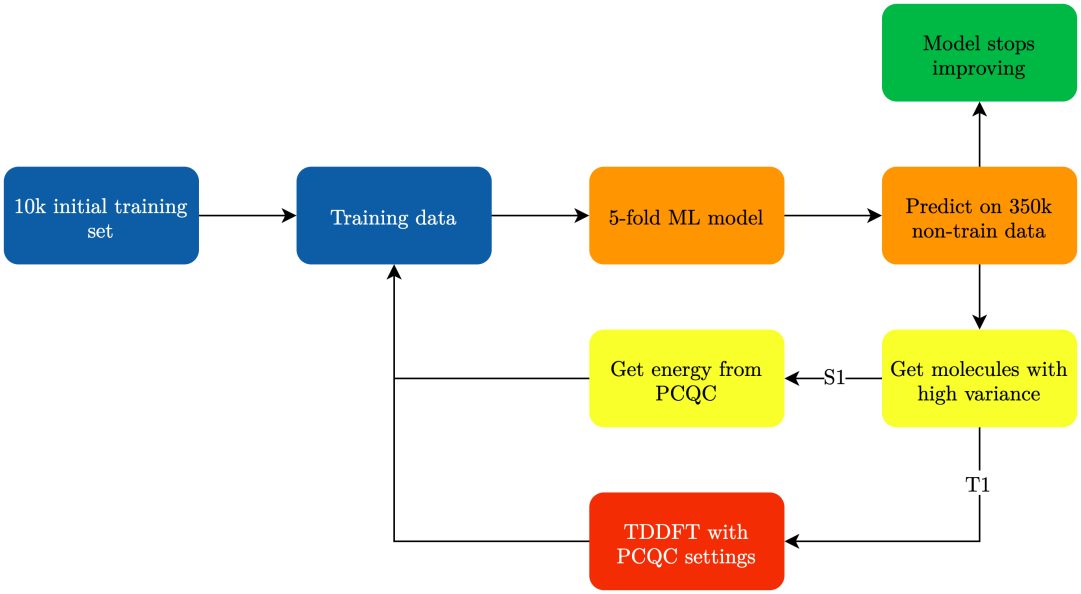


FIGURE 3.2: Active learning workflow. Initial training set composed of molecules from literature scraping. In each AL cycle, a 10-fold ML model is generated with the training data. The model is then used to measure uncertainty for the remaining molecules in the dataset. Molecules with high uncertainty are extracted, labeled, and added to the training set. Once the number of molecules with high uncertainty is low enough and the MAE on the test set has stabilized, we stop AL cycles and obtain a final training set. Blue boxes represent data, yellow boxes represent quick steps, orange boxes represent moderately slow calculations, and red boxes represent very slow operations. Green box represents final model.

Starting with a small initial training set (10k molecules generated from literature scraping of relevant molecules, detailed in Section 4.2.2), an ML model is generated and used to predict on the larger database. As before, a 10-fold ML model is used as an ensemble, but instead of averaging the results, here an epistemic uncertainty value is calculated using the variance in the models. Specifically, the uncertainty for molecule i is defined as:

$$\rho_i = \frac{\sigma_i}{\sqrt{N_i}} \tag{3.1}$$

where σ_i is the ensemble variance and N_i is the number of atoms for molecule i. This expression is derived from Smith et al. (2018).[139] They divide by $\sqrt{N}$ because they are trying to predict energy - although there is no physical motivation behind dividing by $\sqrt{N}$ for excited state energy, it helps improve stability. Uncertainty is used as a predictor of error, where error is defined as:

$$\varepsilon_i = \frac{|\max(E_i^{ens_k} - E_i^{ref})|}{\sqrt{N_i}} \tag{3.2}$$

where $E_i^{ens_k}$ is the energy for ensemble k, E_i^{ref} is the reference energy from the database for molecule i. Again, this definition is derived from Smith et al. (2018)[139] and is used for stability of the AL cycles. As seen in Figure 3.3, there is a general correlation between uncertainty and error as defined above.

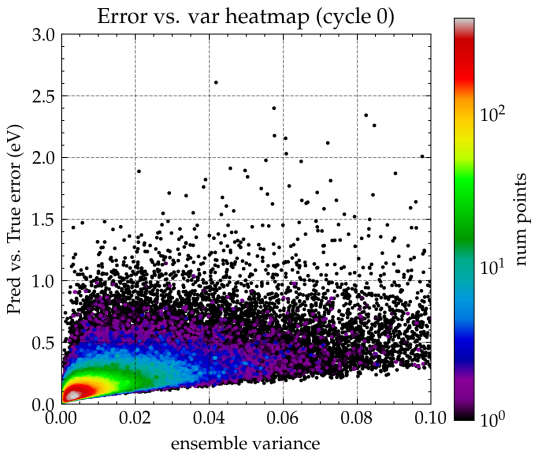


FIGURE 3.3: Heatmap of scatter plot of error vs. uncertainty for S1 energies of molecules predicted with 10-fold ML model trained on the initial training set. Shows general correlation between uncertainty and error. Compare with Figure 1 of Smith et al. (2018).[139] Note that most molecules are concentrated at low error and low uncertainty, while there is a general positive correlation between uncertainty and error.

Once the uncertainty for molecules is calculated, high-uncertainty molecules are chosen to be labeled, and depending on the energy type desired (S1 or T1) the data is either extracted from the database (for S1) or calculated with TDDFT (T1). The labeled molecules are then added to the training set, and the AL cycles continue.

This AL workflow can be optimized by tuning certain parameters. Two forms of optimization are done here, first with the initial training set, and second with the rules for molecule addition per cycle.

3.2.2 Initial training set optimization

The initial dataset of 10k molecules is composed of molecules scraped from literature abstracts tagged with "triplet-triplet annihilation" or "singlet fission" (more details provided in Section 4.2.2). However, due to the nature of this scraping, it is possible that many molecules are fairly similar, creating redundancies in the data,

which limits the wider applicability of the initial training set to general molecules. One way to reduce redundancy is by removing molecules whose properties can be predicted by a ML model trained on other molecules in the dataset. This training set optimization is inspired by Figure 2(a) in Smith et al. (2018),[139] adapted for this work as shown in Figure 3.4.

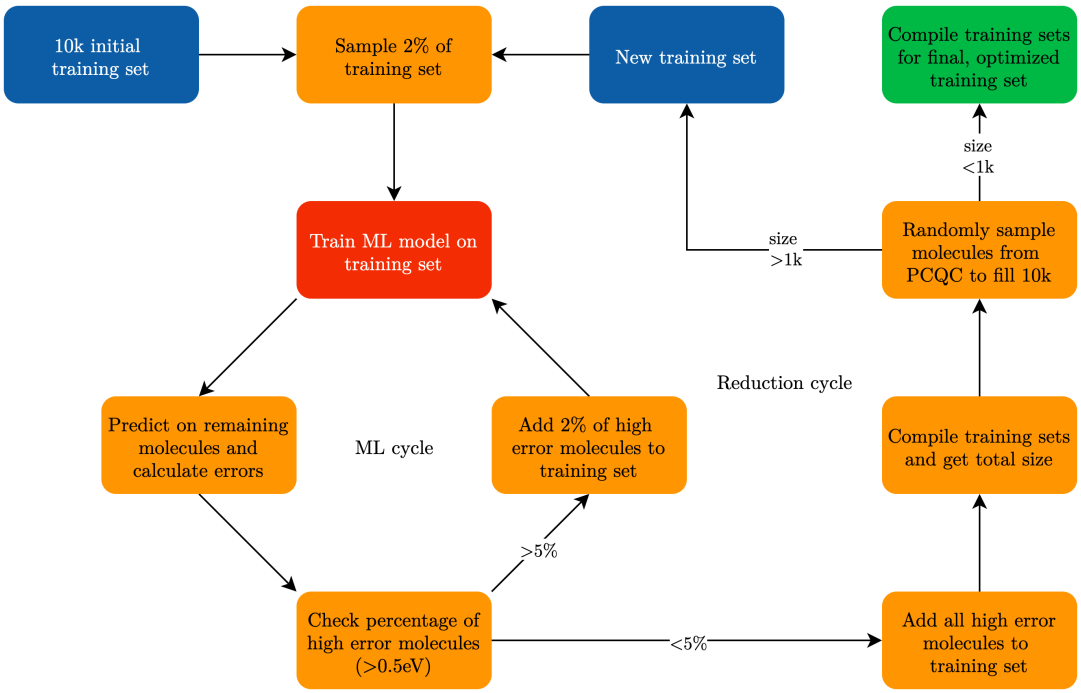


FIGURE 3.4: Initial dataset optimization workflow. Blue boxes represent data, orange boxes represent quick steps, red boxes represent slow operations, and green box represents final optimized training set.

The workflow starts by randomly sampling 2% of molecules from the initial training set. Then, an ML model is generated and used to predict properties of the remaining non-training data, and the error is calculated between predicted values and reference values (for S1, in the PCQC database, and for T1, calculated with TD-DFT). Then, if more than 5% of the remaining molecules have high error (defined as > 0.5 eV), then 2% of the high-error moleucles are added to the training set, and the ML cycle repeats. If less than 5% of the remaining molecules have high error, then all of the high error molecules are added to the training set. This is then defined as the reduced training set. The reduced set is compiled with any previous reduced sets, and the size of the total training set is calculated. Then, molecules are randomly sampled from PCQC to increase the size of the total training set to 10k. If greater than 1k molecules are added, then the reduction cycle restarts with those sampled molecules to generate a new reduced set. If less than 1k molecules are added, then these are considered the last reduced set, and are compiled with all previous reduced sets to form the final, optimized training set.

Figure 3.5(a) shows the ML cycle round 1, starting with the 10k initial training

set and ending with a reduced 3.5k set. The dataset gradually grows over 100 iterations to ensure all molecules in the reduced training set are essential and non-redundant. Figure 3.5(b) shows the results of the final, optimized training set (labeled 'AL') compared with a randomly selected 10k training set ('RS') and the initial literature-scraped training set (labeled 'SCOP'). The plot shows RMSE of ML predictions on a 50k test set randomly selected from PCQC, and shows the improvement in accuracy after the initial dataset optimization is complete.

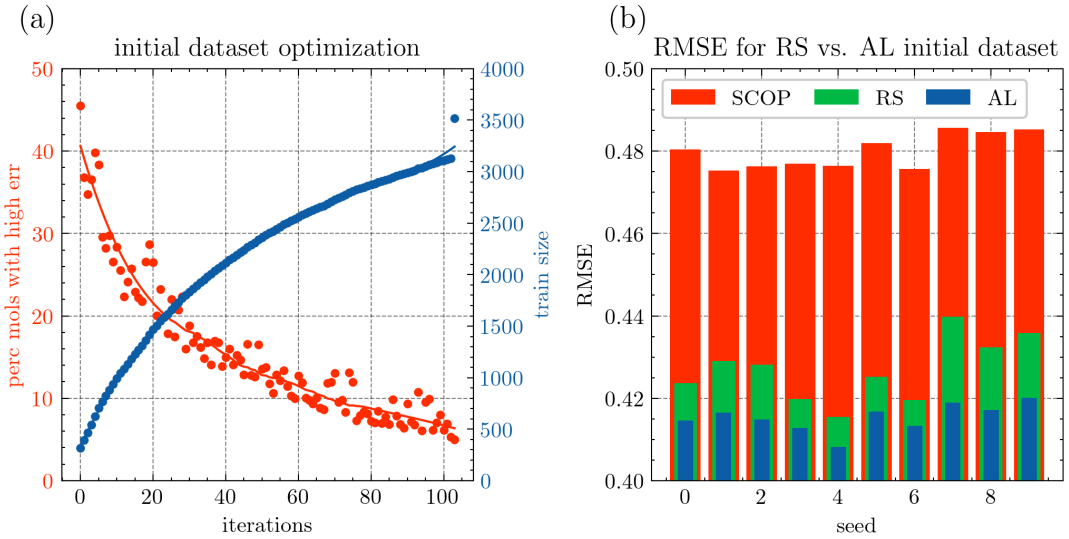


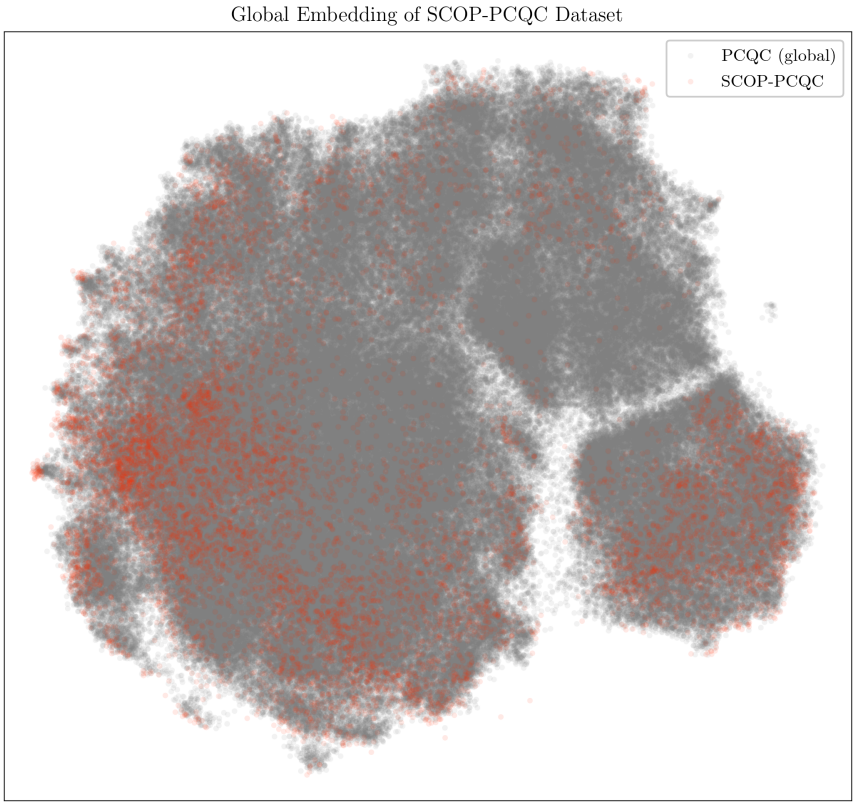
FIGURE 3.5: Results of initial dataset optimization. (a) Shows an example of an ML cycle starting with 10k molecules and ending with a 3.5k reduced dataset, and (b) shows a comparison of the final optimized training set (blue) with a randomly generated dataset (green) and the initial training set (red). The x-axis indicates the seed used in the random generator to create the random training set as well as the 50k test set, to show consistency of results across random sets.

The initial dataset has fairly high RMSE, as expected, likely due to the homogeneity of the data. In contrast, randomly sampled training sets have lower RMSE, but the optimized training set has the lowest RMSE, which validates the optimization approach.

To qualitatively evaluate the improvement, it is possible to visualize coverage of global chemical space by the optimized dataset versus the original. UMAP[141] was chosen for global embedding due to its speed compared to other embeddings such as t-SNE. UMAP embeds high-dimensional molecular data into 2 dimensions, using the Jaccard-Tanimoto similarity between Morgan fingerprints of molecules for proximity. Figure 3.6a shows the global embedding of the literature-scraped dataset (labeled 'SCOP-PCQC') in the global chemical space (labeled 'PCQC (global)'). Similarly, Figure 3.6b shows the global embedding of the optimized initial dataset (labeled 'AL-opt').

As seen, the SCOP-PCQC dataset lacks coverage in certain areas, while clustering in other areas. While this can be useful for predicting properties of certain molecules, it limits the broader applicability of this dataset. In contrast, the AL-opt

(a)



(b)

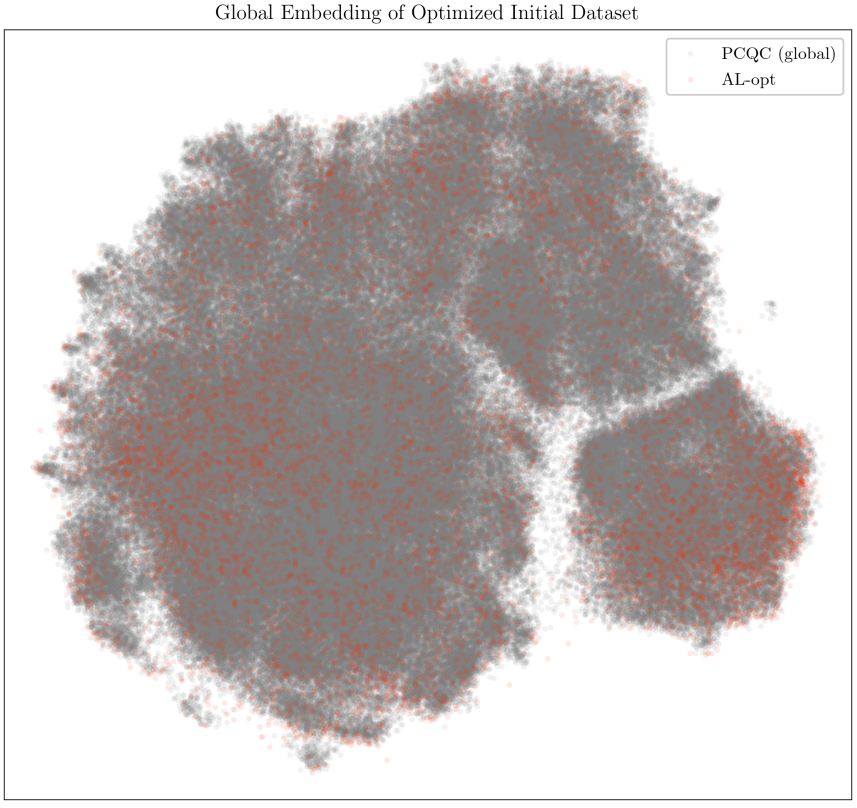


FIGURE 3.6: Global embedding of (a) literature scraped and (b) optimized training datasets in PCQC. UMAP was used for global embedding of PCQC (grey), and the model was used to predict locations of the specific datasets (red).

dataset has much broader coverage, with less clustering of datapoints. Although the distribution may look random, it indeed performs better than a randomly generated training set, as proven in Figure 3.5. This is likely because the optimized dataset is sparse in areas where molecules are similar and therefore easy to predict, and more condensed where molecules are different.

Now that an optimized initial training set has been generated, we can move onto optimizing the molecular additions for each AL cycle.

3.2.3 AL cycle molecule additions

The next important optimization is choosing which molecules to add to the training set for each AL cycle. In Figure 3.3, while there is a general correlation between high uncertainty and high error, there is not a clear correlation. It is therefore necessary to decide thresholds for "high error" molecules and "high uncertainty" molecules. Figure 3.7(a) shows a colormap of the distribution of possible error thresholds and uncertainty thresholds, and the percentage of high-error molecules included for each point on the grid. Figure 3.7(b) shows a cross-section of (a), with error threshold fixed at 0.3 eV. Uncertainty and error are defined in Equations 3.1 and 3.2 above.

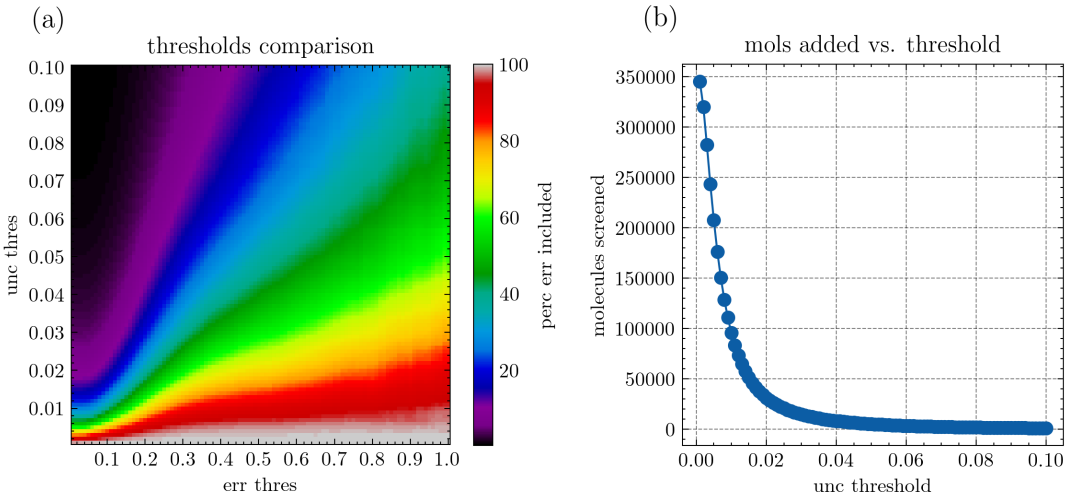


FIGURE 3.7: (a) Plot of percentages of high-error molecules included as a function of error and uncertainty thresholds. (b) Plot of total number of molecules added as a function of uncertainty threshold, assuming an error threshold of 0.3 eV.

As seen, depending on the error threshold used to define molecules as "high error", it can be easy or very difficult to capture a large fraction of "high error" molecules. Because the objective is to generate a highly accurate ML model, a strict error threshold of 0.3 eV is chosen. One drawback of having a low error threshold is the large number of molecules added per AL cycle - as seen in Figure 3.7(b), the number of added molecules exponentially increases as the uncertainty threshold decreases. Figure 3.8 presents two plots to help understand the distributions of total molecules added versus high-error molecules added, assuming an error threshold of 0.3 eV.

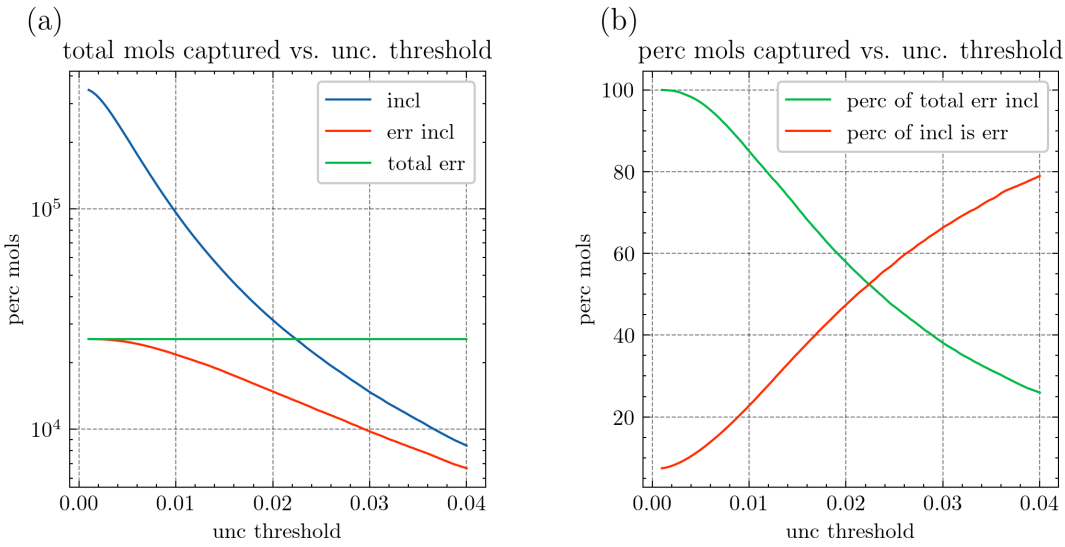


FIGURE 3.8: (a) Plots of number of molecules captured as a function of uncertainty threshold for a fixed error threshold of 0.3 eV. Green line shows total number of high-error molecules, while red line shows number of high-error molecules included. Blue line shows total number of molecules included. (b) Plots of percentage of high-error molecules captured. Green line shows percentage of high-error molecules included in captured data, while red line shows percentage of captured data that is high-error.

Figure 3.8(a)'s red and green lines show the benefits of a low uncertainty threshold. The green line shows the total high-error molecules, and the red line shows the number of high-error molecules included as a function of uncertainty threshold. As seen, a lower uncertainty threshold increases the number of high-error molecules added. Unfortunately, it also increases the number of total molecules added (seen in the blue line), expanding the dataset. To explore this further, Figure 3.8(b) shows the percentage of high-error molecules captured, as well as the percentage of captured molecules that are high-error. As seen, as the uncertainty threshold decreases, the percentage of high-error molecules included increases, but because so many extraneous molecules are added, the high-error molecules make up a smaller percentage of the total added dataset.

Because the primary objective of this workflow is increasing the accuracy of the ML model, a low uncertainty threshold is tolerable, despite the large number of molecules added per cycle. 0.01 was chosen as the uncertainty threshold, corresponding to 85% of high-error molecules being included in the added dataset, and 100k total molecules added. Note that as the AL cycles continue and the model improves, fewer molecules should be labeled as high-error.

Now that the AL workflow has been optimized, we can begin running AL cycles and analyzing the results.

3.3 Results

3.3.1 Singlet AL

The AL workflow was run for 8 cycles, and the training sets at each cycle were tested on a randomly generated 350k molecule test set. As before, an ensemble ML model was trained on 10 folds, and the predictions were either averaged to give the final prediction, or used to calculate the epistemic uncertainty as defined above. The test sets were pruned to avoid any molecules in the training set. Figure 3.9 shows the performance of the ML model at each AL cycle, as well as the training set size.

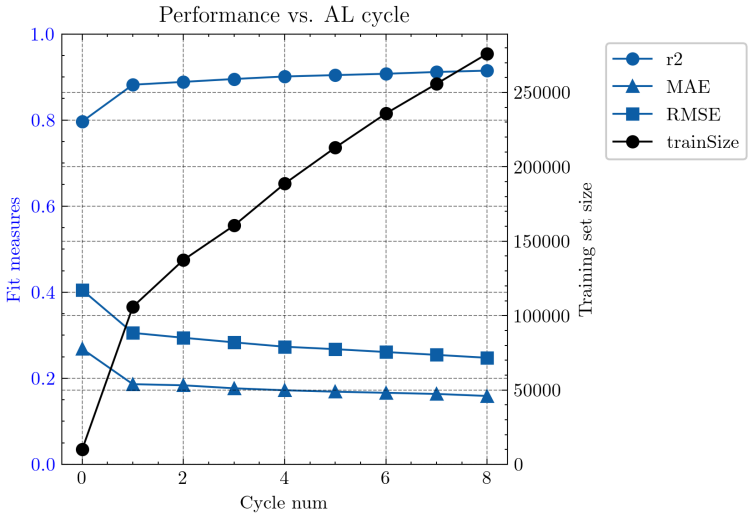


FIGURE 3.9: Plots of various performance measures of the ML models generated at each cycle of AL. ML models were trained on the training set comprised of the previous cycle's molecules plus the added molecules from uncertainty analysis. They were then used to predict S1 energies on a randomly generated 350k molecule test set, and the predictions were compared against the ground truths in the PCQC database.

As seen, the R2 consistently increases, and MAE/RMSE consistently decrease, with each additional AL cycle. The largest improvement is with cycle 1, which adds around 100k molecules to the training set. After cycle 1, gradual but consistent improvements can still be seen.

To show that the uncertainty and error thresholds defined in the previous section help improve the ML performance, Figure 3.10 shows plots of uncertainty versus error for the beginning of AL (cycle 0) and the end (cycle 8). As seen, the percentage of predictions that are either high error or high uncertainty drastically decrease through AL, with only 1.35% of molecules exhibiting high error and 6.27% of molecules exhibiting high uncertainty at the last AL cycle.

We can also visually see the improvement at cycle 8 by plotting the ML predictions against the database reference values, as shown in Figure 3.11. As seen, there are fewer outliers, and more points are located on the $x = y$ line. Quantitatively, the R2 increases by 0.11 points and the MAE/RMSE decrease by 40%.

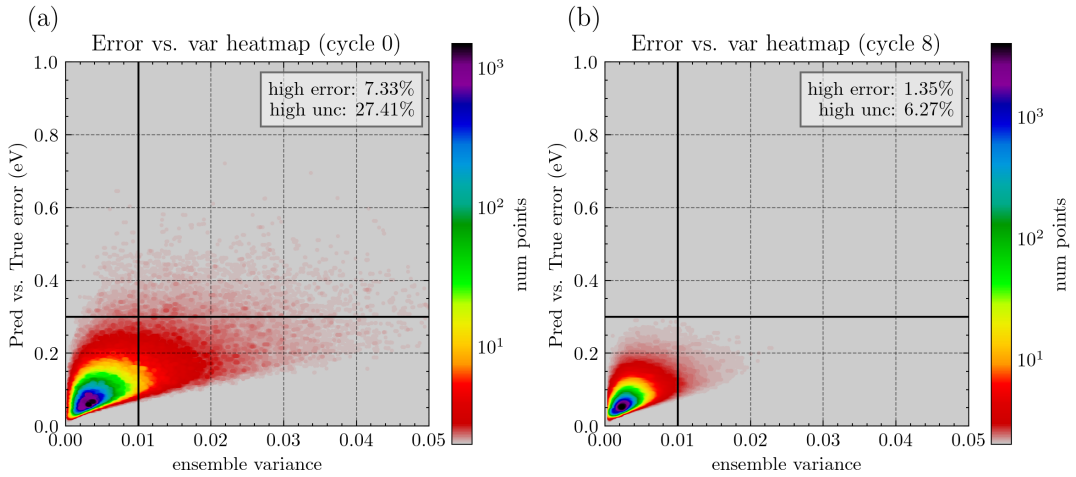


FIGURE 3.10: Plots of heatmaps of error vs. uncertainty for (a) cycle 0 (reproduced from Figure 3.3 for convenience), and (b) cycle 8. Black lines show the error and uncertainty thresholds used in this workflow. Inlaid data shows quantitative measurements of improvement of ML model by reducing high-error and high-uncertainty predictions.

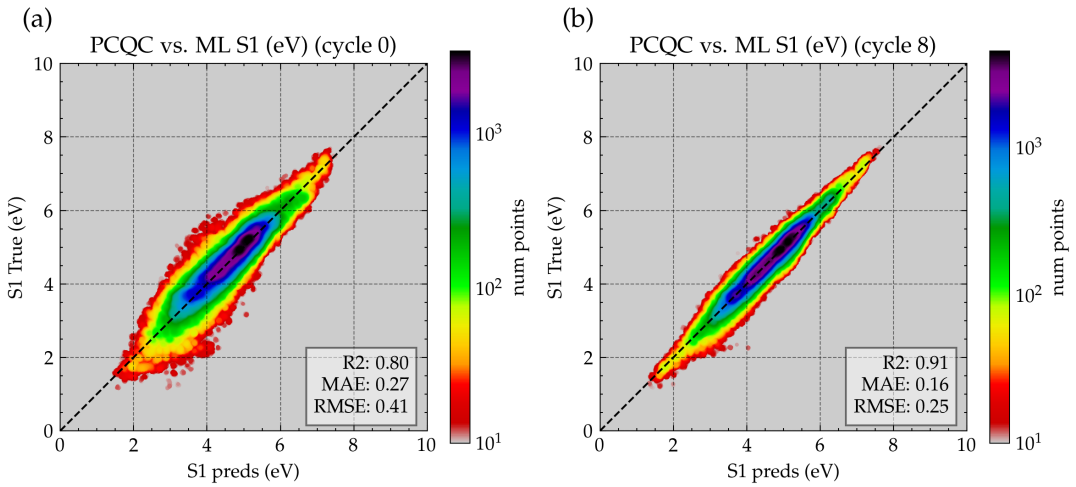


FIGURE 3.11: Plots of heatmaps of predictions vs. reference for (a) cycle 0 and (b) cycle 8. Dashed black line shows the $x = y$ line. Inlaid data shows quantitative measurements of improvement of ML model.

To qualitatively evaluate the AL cycles, we can plot the added molecules in global chemical space. Figure 3.12a shows plots of molecules added in each AL cycle (colored according to AL cycle) in global chemical space, Figure 3.12b shows contour density plots of molecules added per cycle, while Figure 3.12c shows a density plot of the final training set in global chemical space. Again, UMAP was used with 350k molecules randomly sampled from PCQC as the global reference, with positions of the training sets predicted accordingly.

As seen in these plots, the AL workflow allows adaptive training set generation. In Figure 3.12a, all cycles tend to broadly cover the global chemical space, but as seen from Figure 3.12b, even starting from cycle 0 additions, the AL workflow is able to identify high-uncertainty areas. As the cycles progress, the AL workflow

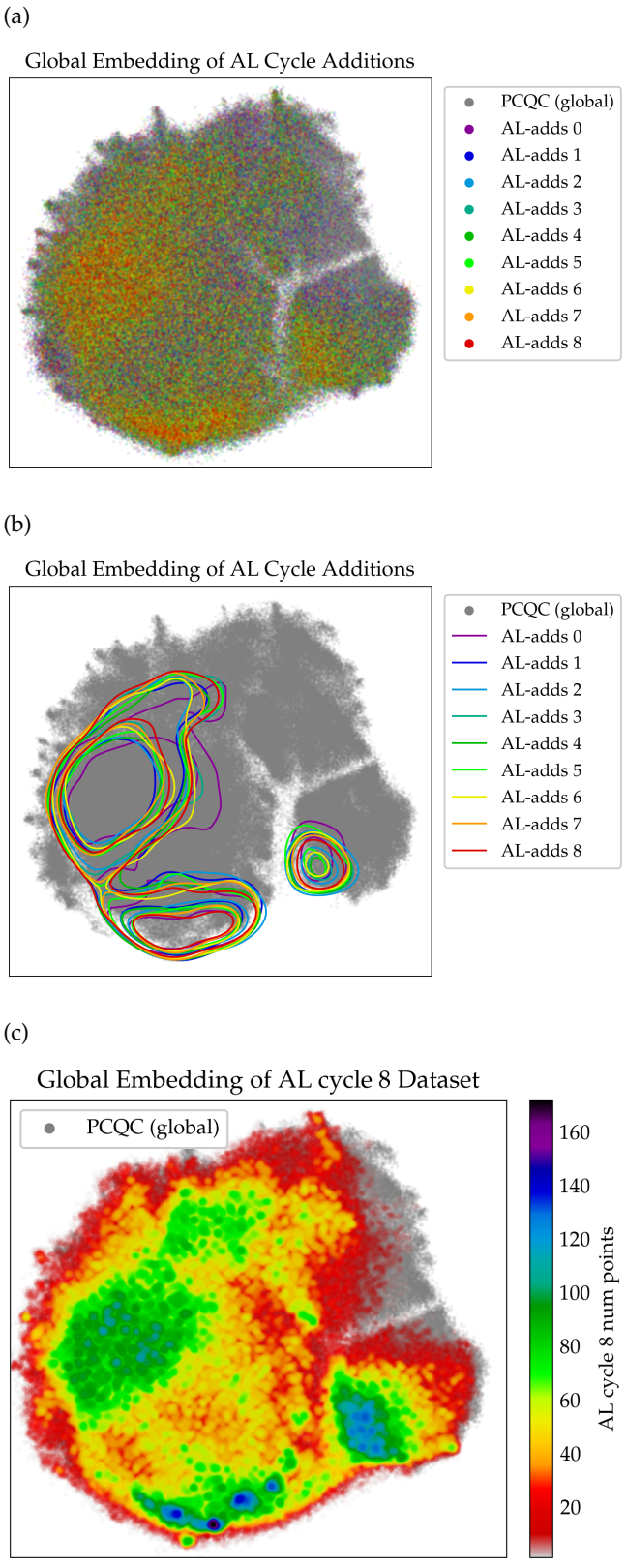


FIGURE 3.12: Global embedding of (a,b) AL added molecules and (c) final training set (cycle 8) in PCQC. (a) Shows all molecules while (b) shows a contour plot for clarity. The contour plot splits the data into 3 sections, 50% of molecules are below the outer line while 25% are above the inner line. (c) Shows a density plot of number of molecules in space. Density of points for (b) and (c) were calculated using Gaussian kernel-density estimation. UMAP was used for global embedding of PCQC (grey), and the model was used to predict locations of the specific datasets (colored).

continues to sample the high-uncertainty areas, as seen by the red and yellow datapoints/contour lines. Specifically, the left center region, lower region, and right lower region have a higher concentration of datapoints, indicating these regions are over-sampled by AL. This over-sampling is more clearly seen in 3.12c, where higher concentrations of molecules are located in the same sections of chemical space as described before. This indicates that even though these molecules are nearby in chemical space, their excited state properties may be quite different and difficult to predict, as small changes in chemical structure could create large differences in S1 energy. In contrast, other regions of chemical space are more sparsely covered by the AL set, indicating properties of these similar molecules are more easily predicted.

From these plots and data, it is evident that the AL cycles produce a viable, accurate ML model by building up an optimized training set. To prove the performance increase is not simply a result of a larger training set, but rather is due to intelligent training set construction with AL, we can compare the AL results to random sampling.

3.3.2 Comparison to random sampling

We first compare the performance of the final ML model from AL to the conventional ML model in Section 3.1.1, formed from randomly sampled (RS) molecules. Table 3.1 shows this comparison.

TABLE 3.1: Performance of final AL model vs. RS model

	RS	**AL**
R2	0.896	0.915
MAE	0.153	0.159
RMSE	0.287	0.248
Train Size	500,000	276,013

As seen, the AL model outperforms the RS model in virtually all measures, with only a very slightly higher MAE. This is despite the training size of the AL model being 55% the size of the RS model, creating a significant time savings for calculations. This shows the importance of intelligently formulating the training set instead of relying on random sampling of the chemical space.

To further prove this point, at each cycle of AL, an equivalently sized training set composed of randomly sampled molecules was generated. An ML model was then trained on this data and tested on the same test set as the AL data. Figure 3.13 shows a comparison of RMSE for ML models trained on RS versus AL training sets. As seen, the AL models outperform the RS models, slightly at first but more

strongly as the cycles increased. The RS models start to stagnate in performance improvements, while the AL models continue to steadily improve.

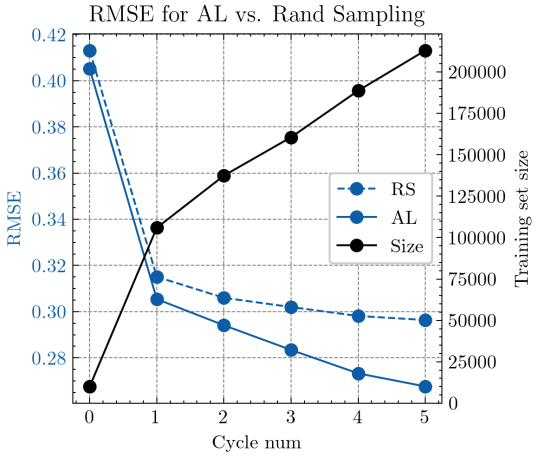


FIGURE 3.13: Plot of RMSEs of ML models trained on actively learned (AL) or randomly sampled (RS) data for each AL cycle. RS data is composed of randomly sampled molecules of the same size as the AL training set. Both ML models are tested on 350k randomly sampled molecules, pruned for no overlap with either training set.

From these analyses, it is clear the AL model is able to predict S1 energies accurately after 8 cycles. The training set size is only half of what would be needed with random sampling of molecules. The next step is to apply this AL workflow to triplet energies, and analyze the results.

3.3.3 Triplet AL

The same AL workflow is applied to triplet energies to generate an accurate ML model with a minimal training set. The initial dataset is the same as described in Section 3.2.2, and the same uncertainty threshold of 0.01 (as defined in Section 3.2.3) is used. At the time of writing, only 1 AL cycle with T1 energies was able to be completed. 161,710 molecules were labeled as high-uncertainty after the AL cycle. Figure 3.14 shows a histogram of uncertainties, for reference. Of the 162k added molecules, T1 energies of 133,186 were successfully calculated with B3LYP/6-31+G(d) TD-DFT. Only one cycle was able to be completed because 133k TD-DFT calculations took several weeks to complete.

Adding these 133k molecules to the initial 10k training set gave a new training set size of 143k. Similarly to S1, the molecules were plotted in global chemical space to get a sense of relative distribution and concentrations of each training set, shown in Figure 3.15.

As seen, the initial added molecules match the distribution in Figure 3.6b, with broad coverage and a few areas of relative concentration. The first AL cycle then focuses more on certain areas of high uncertainty, namely the bottom left as well as the bottom left of the rightmost grouping of molecules. This generally matches the distribution of added molecules for S1 (Figure 3.12b), indicating the ML model has

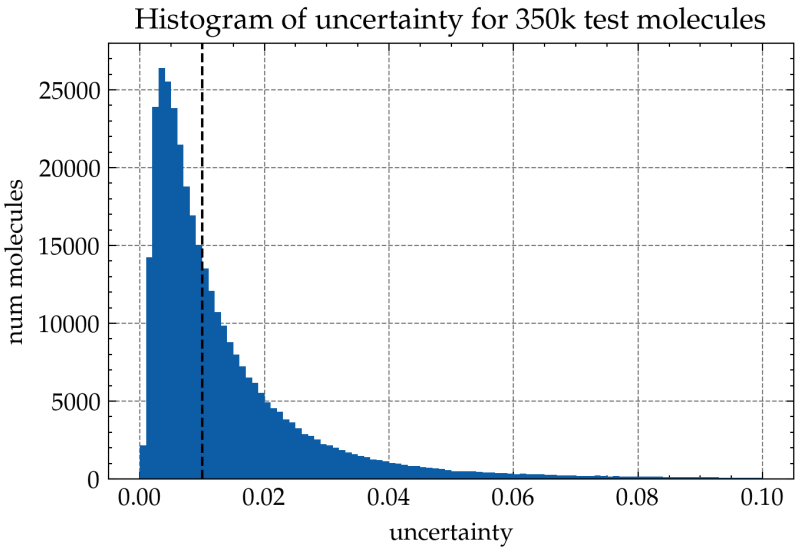


FIGURE 3.14: Histogram of T1 energy uncertainties for 350k test molecules, based on the ensemble ML model generated from the initial 10k training set. Uncertainties calculated based on Equation 3.1. Dashed line shows 0.01 uncertainty threshold.

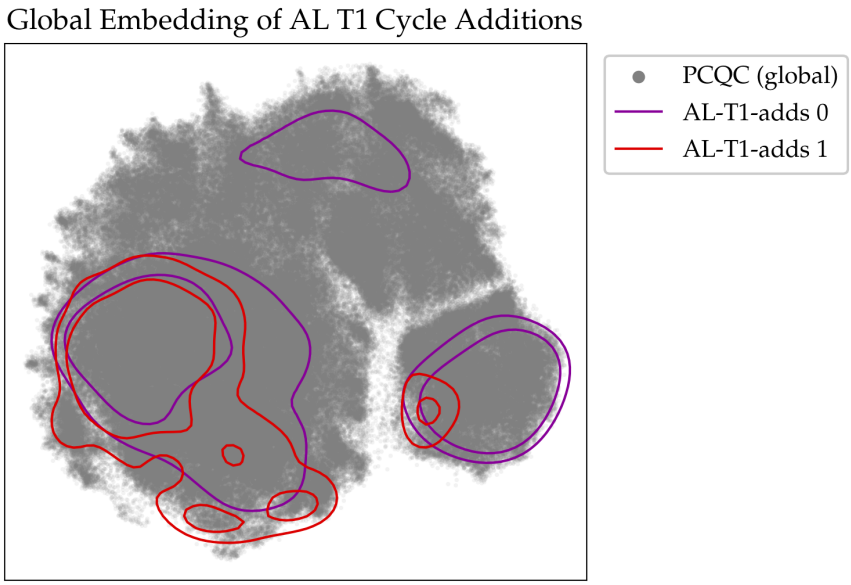


FIGURE 3.15: Global embedding of AL added molecules per T1 cycle. Kernel density estimate (KDE) plot with 50% of molecules lying within the outer circle and 25% of molecules within the inner circle.

trouble predicting S1 and T1 energies in the same regions of chemical space. This makes sense as the same structures influencing the S1 excited state would likely influence the T1 excited state as well.

For S1 energies, the ML models from each AL cycle were able to be tested on a large, 350k test set due to energies being already available in the dataset. For T1, a smaller 10k test set was created of unique molecules not present in the 143k total training molecules. The following plot compares MAE and R2 for the two ML models used to predict properties on the 10k test set.

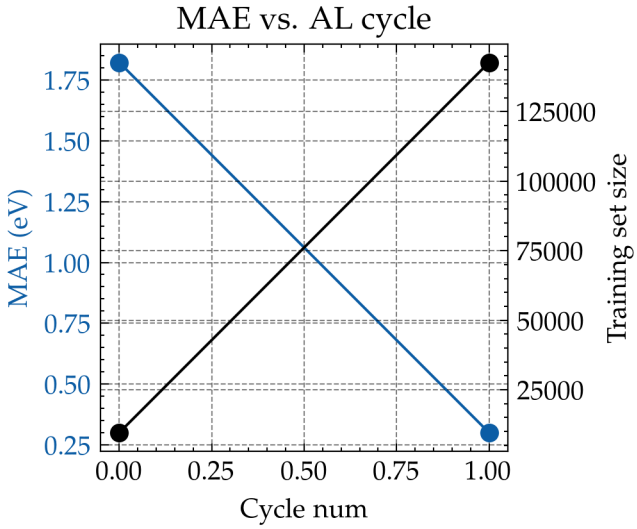


FIGURE 3.16: MAE and training size for the 2 AL training sets completed. Cycle 0 is the baseline with 10k initial molecules, while cycle 1 added 133k molecules. Blue line is MAE and black line is training set size.

Based on this analysis, we can see that the T1 predictions are reasonably accurate (at least for a high-throughput screening technique) with an MAE of approximately 0.3 eV. This is an immense improvement over the initial MAE of 1.8 eV. Unfortunately, a random sampling comparison cannot be done for T1, as this would require twice as many calculations. Thus, we cannot isolate the effects of a larger training dataset size from the implementation of active learning. However, the vast, 6-fold improvement in MAE is promising, serving at least as a proof of the efficacy of AL.

Now that we have accurate ML models for predicting both S1 and T1 energies, we can use the models to identify candidate chromophores within PCQC.

3.3.4 Identifying candidate chromophores

The PCQC database is immense, with 3.5M molecules, and screening molecules with computational chemistry techniques would be prohibitively slow. Instead, we can use the S1-ML and T1-ML models generated in this work to conduct HTVS. We are specifically interested in identifying chromophores for TTA and SF. This entails predicting the suitability of each molecule as a sensitizer (TTA) or emitter (TTA/SF), based on the following suitability functions:

$$\varepsilon_{sens} = e^{-\left|1-\frac{E_{S1}}{E_{T1}}\right|} \tag{3.3}$$

$$\varepsilon_{emit} = e^{-\left|2-\frac{E_{S1}}{E_{T1}}\right|} \tag{3.4}$$

where ε indicates the suitability figure of merit (FOM), "sens" refers to sensitizers, "emit" refers to emitters, E_{S1} is the S1 energy, and E_{T1} is the T1 energy. This definition is useful as it is normalized and therefore indicates the suitability of the energy level alignment on a scale of 0 to 1. The emitter FOM is particularly useful as

it allows simultaneous detection of both types of emitters (TTA and SF), which can be distinguished in post-processing.

The ML models were run on all 3.5M molecules in PCQC to predict S1 and T1 energies, which are available on GitHub.[142] Then, the suitability of each molecule as a sensitizer or emitter was calculated. For post-processing, strict bounds were set for TTA sensitizers, TTA emitters, and SF emitters, as described below:

$$\begin{aligned} \text{sens}_{TTA} &: 1.0 < \frac{S1}{T1} < 1.05 \\ \text{emit}_{TTA} &: 1.9 < \frac{S1}{T1} < 2.0 \\ \text{emit}_{SF} &: 2.0 < \frac{S1}{T1} < 2.1 \end{aligned} \tag{3.5}$$

Applying these strict bounds to the full dataset resulted in 307,216 sensitizers, 2763 TTA emitters, and 1694 SF emitters being identified. The SMILES, predicted S1 and T1 energies, and FOM for all identified chromophores are available on GitHub.[142]

Note that there are far fewer emitters identified than sensitizers. To further explore this phenomenon, a histogram of S1/T1 ratios is presented in Figure 3.17. As seen, most molecules have S1/T1 close to 1. This drops off steeply for ratios less than 1 (as this would be an inverse split and is rare), and gradually declines for ratios greater than 1. To understand this decline, we turn to the theory of singlet-triplet splitting, which states that a higher HOMO-LUMO overlap and smaller spatial separation leads to a higher singlet-triplet split.[143] Most materials will have small HOMO-LUMO overlap, due to the difference in molecular properties characteristic of donors vs. acceptors. This could help explain why there are fewer emitters identified than sensitizers.

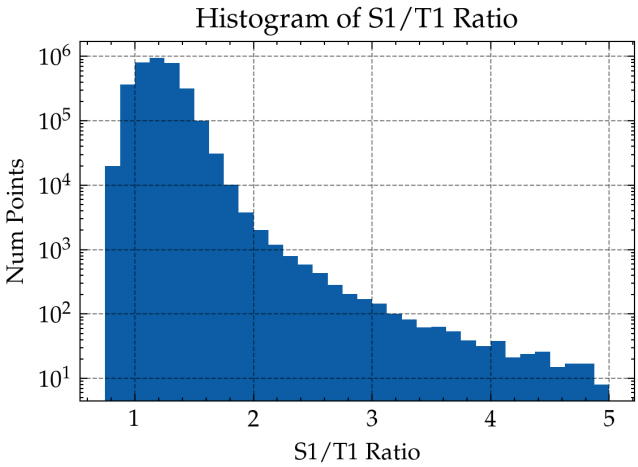


FIGURE 3.17: Histogram of S1/T1 ratios, for ML-predicted S1 and T1 energies.

Of interest to this work are the near-IR (NIR) TTA materials, implying sensitizer S1 between 1 and 1.2 eV, and emitter S1 between 1.9 and 2.4 eV.

Using these strict limits, 56 NIR candidate sensitizers and 243 emitters were identified. Because there were only 300 total molecules identified as potential TTA sensitizers and emitters, it was possible to run TD-DFT to confirm results. Of the 56 sensitizer candidates, 9 were confirmed to be suitable with TD-DFT, and 2 of those were confirmed to lie within the NIR zone of interest. Similarly, of the 243 emitters, 17 were suitable, and 6 operated in the NIR region. The SMILES, predicted S1/T1 energies, predicted FOM, TD-DFT S1/T1, and TD-DFT FOM for these 8 molecules are available in Appendix Table A.1, and the data for all 300 molecules is available on GitHub.[142]

As seen, the model is better at predicting suitable molecules than the exact energies of the molecules. For extra flexibility, the bounds were expanded slightly to 0.7 to 1.5 eV for sensitizer S1 and 1.5 to 2.5 eV for emitter S1. This resulted in 276 near-IR sensitizer and 736 near-IR emitter candidates. These 1000 molecules were then run with TD-DFT to confirm results. Of the 276 sensitizer candidates, 55 were confirmed to be suitable with TD-DFT, and 7 of those were confirmed to lie within the NIR zone of interest. Similarly, of the 736 emitters, 43 were suitable, and 7 operated in the NIR region. These 14 confirmed sensitizers and emitters are presented in Figure 3.18. The SMILES, predicted S1/T1 energies, predicted FOM, TD-DFT S1/T1, and TD-DFT FOM for these 14 molecules are available in Appendix Table A.2, and the data for all 1000 molecules is available on GitHub.[142]

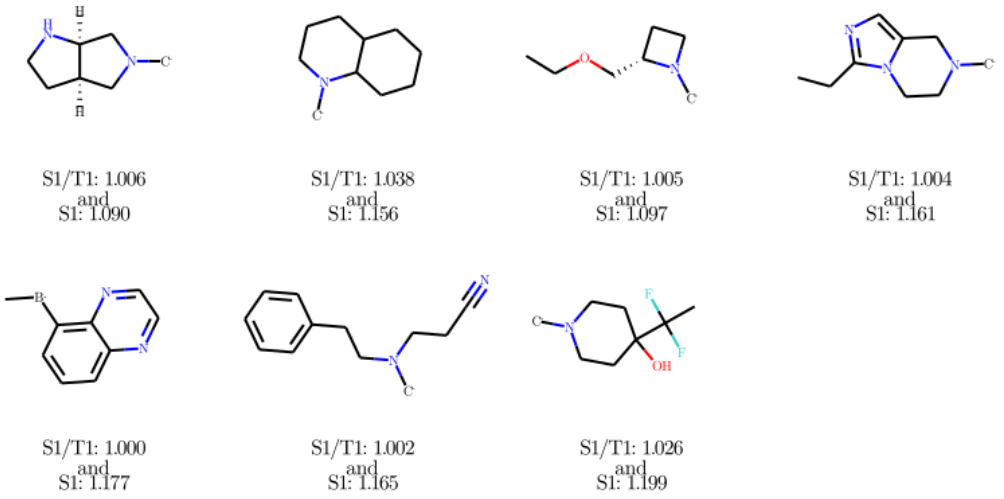
A few differences are evident – emitters are more likely to be aromatic and contain oxygen, while sensitizers generally feature non-aromatic rings and CN atoms. However, because there are relatively few identified molecules, it is difficult to describe general properties of each. It is therefore useful to expand the candidate space, as described in the next section.

Expanding candidate space with GB-GA

Since only a few candidates were identified in the previous section, it would be beneficial to expand the candidate space to identify more potential chromophores. The PCQC database has been exhausted, and while it is possible to turn to other large-scale databases, this is time-consuming and may potentially only lead to a few candidate molecules, as the PCQC database did. To more efficiently generate novel chromophores, we turn to genetic algorithms, specifically the graph-based genetic algorithm (GB-GA) developed by Jensen,[144] featuring crossovers and mutations designed to generate novel "children" molecules from parents.

GB-GA has been shown to perform well in comparison with ML-based methods such as graph convolutional policy networks, with the added benefit of improved computation time of several orders of magnitude.[144] In 2020, GB-GA was expanded to include absorbance calculated with xTB-sTDA.[145] The workflow for absorbance calculations was to generate 20 random conformations with RDKit, use MMFF94 to minimize their energy, and choose the lowest-energy conformer as the input structure.[145] Then, the input structure was directly used with sTDA to get the

(a)



(b)

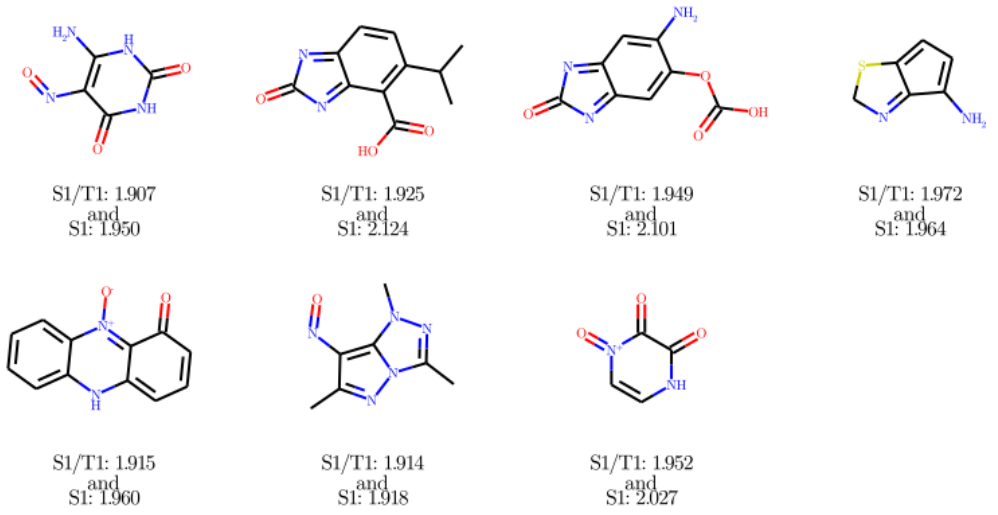


FIGURE 3.18: Candidate (a) sensitizers and (b) emitters predicted by the ML model and confirmed with TD-DFT to have desirable energy level alignment for NIR-to-visible TTA with minimal energy loss.

S1 energy and oscillator strengths, which are summed (after a Gaussian normalization) to get the final molecular score,[145] as shown in Equation 3.6:

$$\text{Score} = \exp\left[-\frac{1}{2}\left(\frac{\lambda - \lambda_t}{\sigma}\right)^2\right] + \frac{\min(\omega, 0.3)}{0.3} \tag{3.6}$$

where λ is the absorbance wavelength of the candidate molecule, λ_t is the target wavelength, σ is the normalization factor, and ω is the oscillator strength. While this is a rapid methodology capable of generating thousands of child molecules per run, it is potentially inaccurate due to the limited ground state optimization conducted. Ideally, xTB would be used for fast ground-state optimization, but it would be too slow for the genetic algorithm. Instead, this study uses the ML models generated in this work. Because the ML models directly predict energies from SMILES, they circumvent the requirements of 3D initialization, conformer searching, and ground state optimization, and can rapidly output energies.

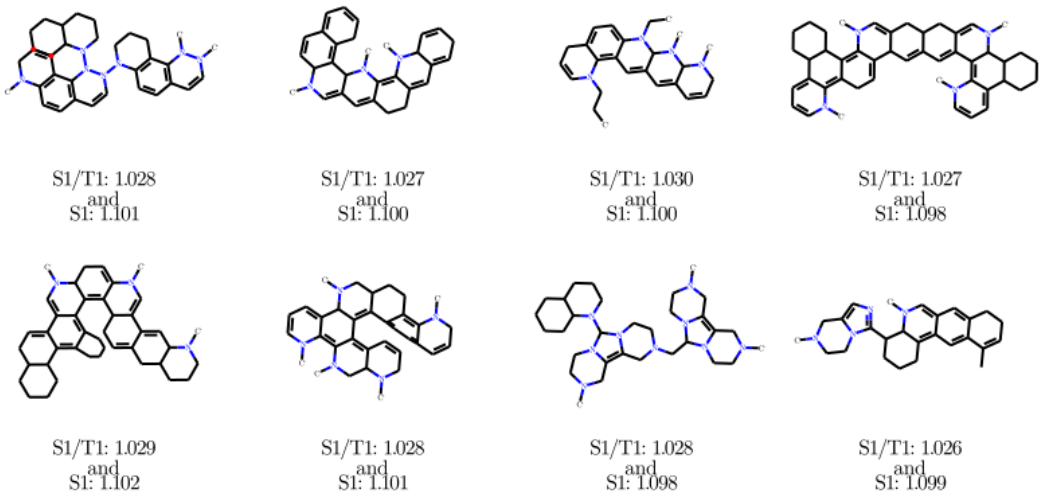
The 7 candidate sensitizers and 7 candidate emitters were used as the initial population pool. For sensitizers, the target S1 was set to 1.1 eV and target T1 to 1.07 eV (97% of S1). For emitters, the target T1 was set to 1.05 eV and the target S1 was set to 2.1 eV. For Gaussian normalization, a σ of 25 nm was used for sensitizers and 50 for emitters. The oscillator strength term in Equation 3.6 was replaced with another wavelength normalization for triplet energy. 50 generations of the genetic algorithm were run, with a mutation rate of 0.05.

After running GB-GA hundreds of times, just over 10,000 candidate sensitizers and 10,000 candidate emitters were generated, where candidates were defined as achieving a score of 1.5 or higher. Because the population pool as small, there were several duplicates – there were a total of 2,193 unique sensitizers and 3,575 emitters generated. All candidate molecules are available on GitHub.[142] The 8 top-scoring sensitizers and emitters are shown in Figure 3.19, to give a sense of the types of structures generated.

As seen, all molecules have precisely-tuned properties, with energies matching the targets with $\pm$0.01 eV accuracy or better. The sensitizer molecules are vastly different from the population pool, while emitters are more similar to the input molecules. This indicates GB-GA not only explores the local chemical space but is capable of traversing significant distances in global chemical space. Suggested sensitizers are larger molecules often with long chains of 6-C rings, while emitters are more likely to have short chains and 5-C rings. Several emitters also have oxygen atoms, while sensitizers primarily have carbon and nitrogen. While these are the highest-scoring molecules, there were several molecules with high scores, with 322 sensitizers and 835 emitters having scores higher than 1.9.

To investigate the other suggested molecules and determine the general structures suggested by GB-GA to be preferable, it is useful to conduct a scaffold analysis. Scaffolds are essentially the bulk structure of the molecule, without including any functional groups. Scaffold analysis was done with RDKit's MurckoScaffold

(a)



(b)

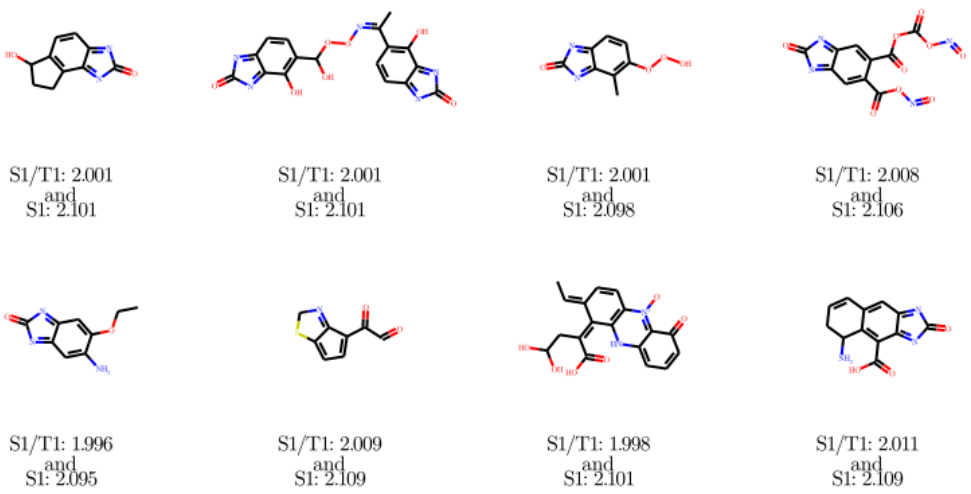


FIGURE 3.19: Candidate (a) sensitizers and (b) emitters generated with GB-GA to have desirable energy level alignment for NIR-to-visible TTA with minimal energy loss. Sensitizers targets were S1 = 1.1 eV and T1 = 1.07 eV, while emitter targets were S1 = 2.1 eV and T1 = 1.05 eV. Energies in eV.

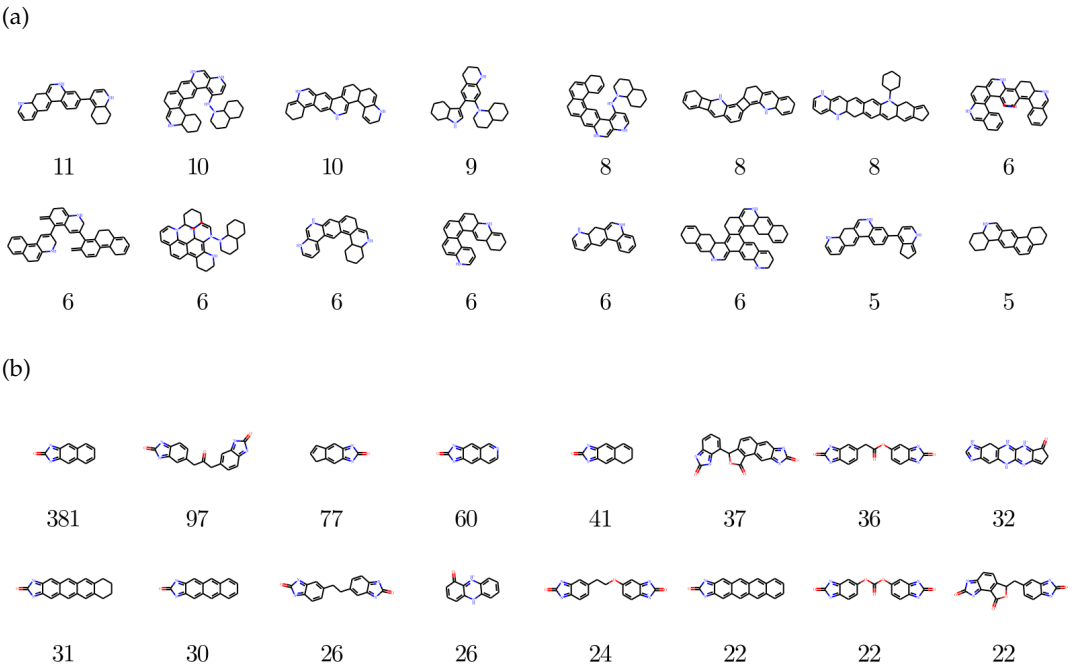


FIGURE 3.20: 16 most common scaffolds for (a) sensitizers and (b) emitters generated with GB-GA, to show general structures preferred by the algorithm. Only scaffolds with more than 10 atoms are shown here.

module. Figure 3.20 shows the 16 most common scaffolds for sensitizers and emitters. As seen, the overarching patterns described above for sensitizers and emitters are upheld, i.e. the long 6-C ring chains in sensitizers and smaller emitter molecules often with 5-C chains. Further, note here the larger diversity of sensitizer molecules, i.e. the most common scaffold appears only 11 times while the most common emitter scaffold appears 381 times. This could suggest a wider variety of sensitizer geometries exist, while emitters may be confined to certain areas of chemical space.

Unfortunately, the generated molecules were not confirmed with TD-DFT. This was due to computational expense – because the molecules are novel and therefore are not in PCQC, their ground-state structure is unknown. This would make TD-DFT confirmation prohibitively expensive. Regardless, based on the analysis in the previous section, a large fraction of these molecules should be suitable chromophores.

3.3.5 Limitations of direct ML

While the ML model overall has high accuracy, the accuracy suffers for low-energy molecules. This is depicted in Figure 3.21, which shows both (a) the MAE per energy interval and (b) the total number of molecules in that energy interval, for S1 energies using a 350k test set, and Figure 3.22, which also shows MAE and number of molecules, but for T1 energies using a 10k test set.

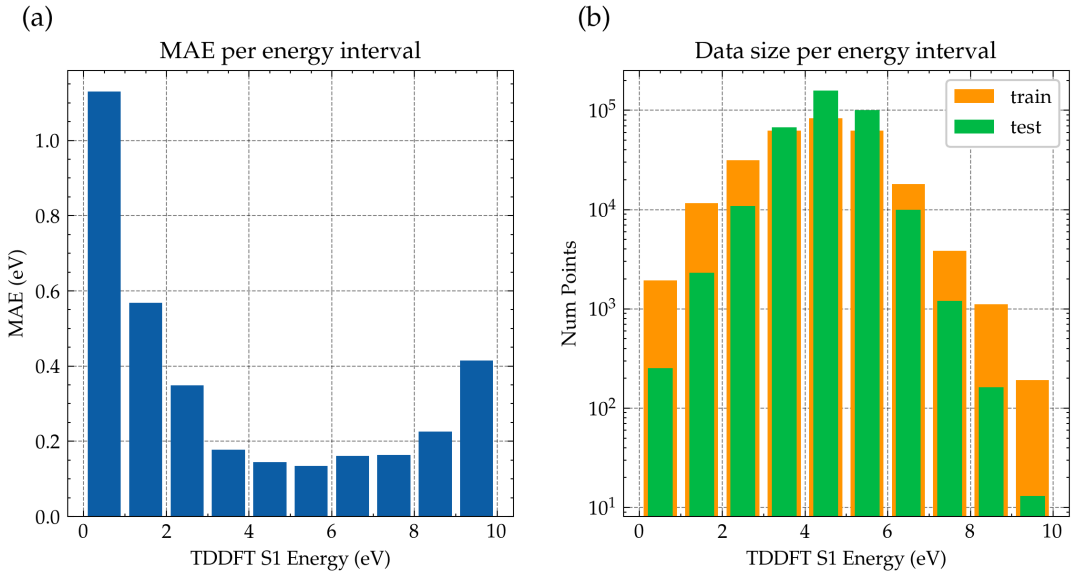


FIGURE 3.21: (a) MAE for all 10 1 eV energy intervals between 0 and 10 eV. (b) Number of molecules for each energy interval, for both the training and test sets. ML model trained on 276k S1 datapoints, and tested on a randomly sampled 350k non-overlapping test set in PCQC.

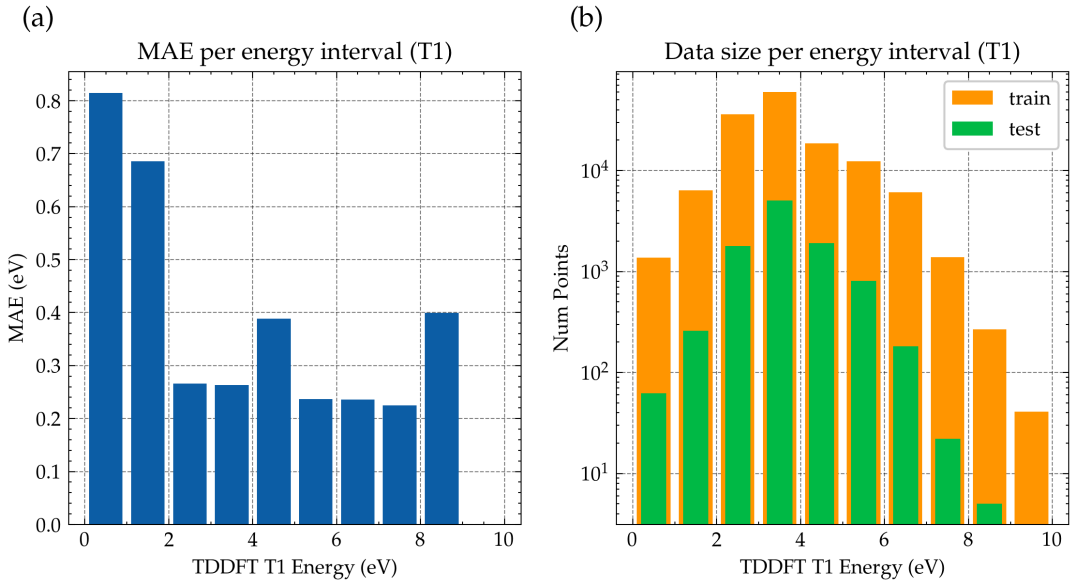


FIGURE 3.22: (a) MAE for all 10 1 eV energy intervals between 0 and 10 eV. (b) Number of molecules for each energy interval, for both the training and test sets. ML model trained on 143k T1 datapoints, and tested on a randomly sampled 10k non-overlapping test set in PCQC.

As seen, while the MAE is low for many mid-energy molecules, it increases drastically as energy decreases. This is likely because comparatively fewer molecules (1-2 orders of magnitude) are in these energy intervals, so the ML model is not able to learn as much about them. However, despite the MAE being high for the absolute energy values, it is still able to predict ratios accurately, as seen in Figure 3.23.

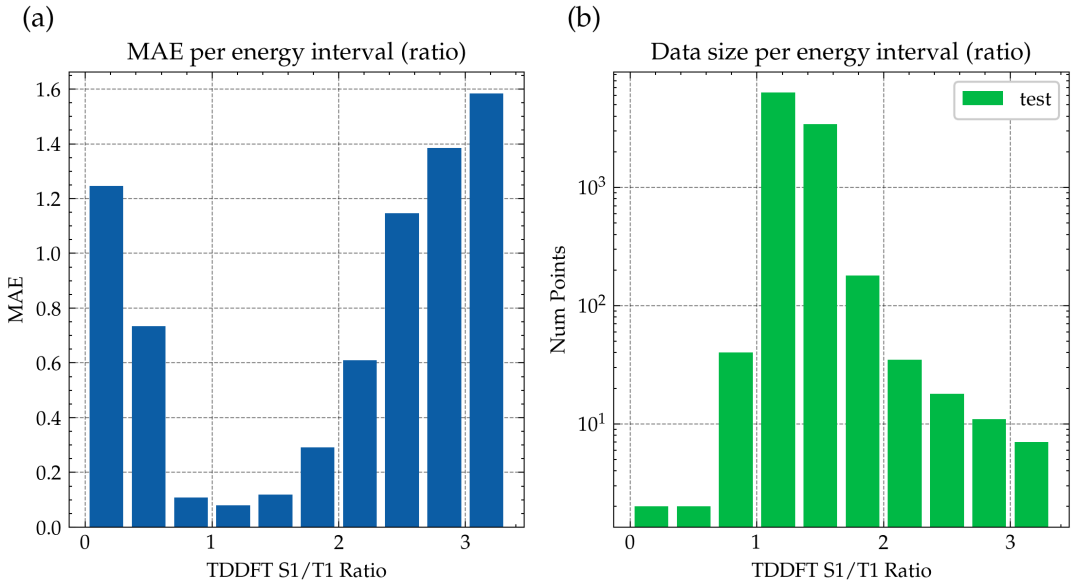


FIGURE 3.23: (a) MAE for all 10 0.33 S1/T1 intervals between 0 and 3. (b) Number of molecules for each energy interval, for the test set. 2 ML models trained separately on 276k S1 datapoints and 143k T1 datapoints were tested on the randomly sampled 10k test set presented previously.

From this figure, it is evident that the ML models are able to predict ratios between 1-2 with good accuracy, while ratios outside of this region may be less accurate. This is again likely explained by the lack of datapoints in the outer regions, with 1-2 orders of magnitude fewer points. Therefore, while the specific energies of molecules predicted in this work may be slightly off, the ratios, and therefore their suitability as sensitizers and emitters, should be accurate. Regardless, a different AL workflow that prioritizes equal distribution of excited state energies could predict absolute energies better.

The following section summarizes this chapter and provides some avenues for future work.

3.4 Conclusions and Future Work

In this chapter, we have developed a machine learning model to accurately predict excited state energies with a small, optimized training set. We generate this training set in cycles, using active learning to identify high-uncertainty molecules among the non-training molecules that should be added to the training set in the following cycle. After optimizing the initial training set and the criteria for adding molecules per cycle, we apply the AL workflow to generate ML models for S1 and T1 prediction. After 8 cycles, S1 prediction had an RMSE of 0.248 eV and MAE of 0.16 eV, with a training set of 276k molecules. In comparison, a randomly sampled training set of 500k molecules had a higher RMSE of 0.287 eV. For T1 energies, after 1 cycle the MAE decreased from 1.8 eV to 0.3 eV, with a training size of 143k molecules.

After generating these ML models, it was possible to predict S1 and T1 energies for all 3.5M molecules in PCQC, rapidly (in approximately 18 hours using a workstation with 24 CPUs). We can then identify potential TTA sensitizers, TTA emitters, and SF emitters by screening this large database. We found thousands of candidate chromophores across a wide variety of energies. Focusing on the near-IR region, 1000 molecules were predicted to be suitable, and running TD-DFT confirmed 14 TTA sensitizers and emitters to be suitable. To expand the candidate space of NIR-TTA materials, a graph-based genetic algorithm (GB-GA) was used, taking the 7 sensitizers (or emitters) as the initial population pool and generating unique child molecules over 50 generations. This was repeated hundreds of times, with the 10,000 highest-scoring molecules chosen. Because the initial population pool was small, there were many duplicates, and 2,193 unique sensitizers and 3,574 emitters were generated. All of the above data is available on GitHub.[142]

The identified or generated molecules constitute the main result of this work. While only a few molecules were ultimately identified as potential NIR-TTA molecules, it is likely there were only a few such candidates in the entire PCQC database. This is because, for sensitizers, only a few molecules (~700) have S1 energies between 0.5 and 1.5 eV in the database. To additionally satisfy the T1 requirements would narrow down the pool significantly. Similarly, for emitters, only a few molecules (~1000) have S1/T1 ratios around 2, and to satisfy the NIR T1 requirement would reduce candidates significantly. Having to run TD-DFT on 3.5M molecules to identify 10s of potential NIR-TTA candidates would be wasteful, but the ML approach presented makes the problem tractable.

There are some avenues of future work to mention here. While the 14 identified NIR-TTA molecules from PCQC were confirmed with TD-DFT, the 5.4k GB-GA generated molecules were not able to be due to time constraints. However, due to the presented accuracy of the ML model and the confirmation of a few molecules within the suggested molecule pool, it is likely several of the molecules suggested by GB-GA would be suitable. Regardless, an immediate next step would be to run TD-DFT on the 5.4k generated molecules.

Another immediate avenue of future work that was limited by computation time is continuing the T1 AL cycles. Because only 1 AL cycle was able to be completed at the time of writing, the T1 accuracy is relatively high (0.3 eV) compared to the S1 accuracy (0.16 eV). Further T1 cycles should improve this accuracy.

Once finalized singlet and triplet training sets are generated, it could be possible to expand beyond S1 and T1 to include the first 10 excited states. PCQC already includes S1-10 data, and by setting the `NStates` keyword in Gaussian to 10, T1-10 energies can also be calculated. Then, the ML model could be changed from single-task to multi-task, and tasked to predict all 10 energies simultaneously. Additionally, adding the oscillator strength (OS) would also be useful. While this study is focused on energy level alignment, improving the efficiency of the TTA process is also critical. OS is the probability of absorption or emission, so a high OS is important for

both sensitizers and emitters. OS is also already output by TD-DFT, so would be an easy addition to a multi-task ML model.

Beyond additions to the current workflow, there are also avenues to improve the workflow. As discussed above, it would be beneficial to adjust the acquisition function to ensure more low-energy molecules are included in the training set. This would likely improve the accuracy of the ML model at low energies, where currently the MAE is high. High S1/T1 ratio molecules also exhibit high MAE, so adding these molecules should also improve accuracy in this regime. Therefore, adding a weighted term for predicted energy and energy ratio in the acquisition function should improve the overall performance of the model.

Finally, while this study focused on NIR-TTA molecules, it is also possible to study other energy regions, and would require simple changes in the bounds set for identification or the target energies requested for GB-GA. These molecules could be useful for other applications beyond solar.

There are some inherent limitations with the ML approach to directly predict excited state energies. First and foremost, ML acts as a black box, with a single output of the desired property. Because of the limited output, it is difficult to develop chemical intuition about the accuracy of results, beyond comparing the test molecule to the training set (though even this may not be a great predictor).

Second, the ML model requires a large training set to output results. Overall, in this study, a training set of 276,013 S1 energies and 133,186 T1 energies was required. Typically, conducting 409k TD-DFT calculations would be prohibitively expensive. This was possible in this work due to ground-state data already available in PCQC, but for another dataset this may not be possible.

Therefore, it may be beneficial to instead use ML to calibrate a high-throughput computational chemistry technique against high-accuracy techniques. High-throughput computational chemistry techniques have well-defined methodology, so researchers can better understand how accurate a technique might be in predicting certain properties of a molecule. Further, a calibration model would likely require less training data, as the load on the ML model would be lightened to just having to shift the calculated energy in the right direction. The following chapter explores this idea further.

Chapter 4

Calibrating xTB-sTDA excited state calculations with ML

4.1 Motivation

As discussed in Section 2.1.1, xTB-sTDA is an ultrafast computational chemistry technique for excited state calculations.[105] Due to its speed, traditionally xTB-sTDA has been used as the first step in high-throughput screening workflows, to filter out large amounts of data into a few candidate molecules, for which properties can be calculated using more advanced, time- and resource-intensive computational techniques.

Although it is an extremely fast technique with calculations for most small molecules completed in under a minute, the trade-off of high speed is potentially low accuracy. For example, Grimme and Bannwarth, in their paper introducing xTB-sTDA, compared vertical excitation energies calculated by xTB-sTDA against SCS-CC2/TD-DFT reference values, and found a mean absolute error (MAE) between 0.34-0.48 eV and standard deviation (SD) between 0.44-0.59 eV, depending on the complexity of input structure.[105]

Because of this potentially high error, if xTB-sTDA is used in high-throughput screening, some suitable molecules may be screened out, or, vice versa, some unsuitable molecules may be included in the candidate pool. To address this issue, the accuracy of xTB-sTDA should be increased, by better calibrating its results against either theoretical results from CC2/TD-DFT or experimental values. Some previous works have attempted to do such calibration, as discussed in the next section.

4.1.1 Previous work in xTB calibration

In the initial paper introducing xTB-sTDA, Grimme and Bannwarth identified the need for calibration, since sTDA does not include solvation or excited state relaxation, estimating results to be blue-shifted by 0.2-0.4 eV.[105] When actually comparing excitation spectra output by xTB-sTDA to experimental excitation spectra, they blue-shifted results by 0.4-1.0 eV, depending on the input structure.[105]

However, shifting xTB-sTDA after seeing experimental results is not practical for wide-scale use. Instead, there should be a way to calibrate xTB-sTDA purely

computationally. Wilbraham et al. used a linear calibration technique to calibrate the first singlet excited state energy (S1) output by xTB-sTDA, training a linear calibration model on 143 molecules and applying the model to 250k small aromatic molecules.[112] They were able to reduce the MAE from 0.258 eV for the original calculations to 0.211 eV for the calibrated data.[112]

However, there are a few issues with this calibration. First, the plot for original vs. calibrated data (Figure 4.1) still shows fairly high error for several molecules, suggesting a simple linear calibration is not sufficient for high accuracy. Second, training a calibration model on only 143 molecules may limit the accuracy of the model when applied to larger datasets, for example the 250k small aromatic molecules considered in the paper. Finally, while Wilbraham et al. also calibrate ionization potential (IP) and electron affinity (EA), they do not calibrate further excited state energy data such as triplet states, higher-level singlet states, or oscillator strength.

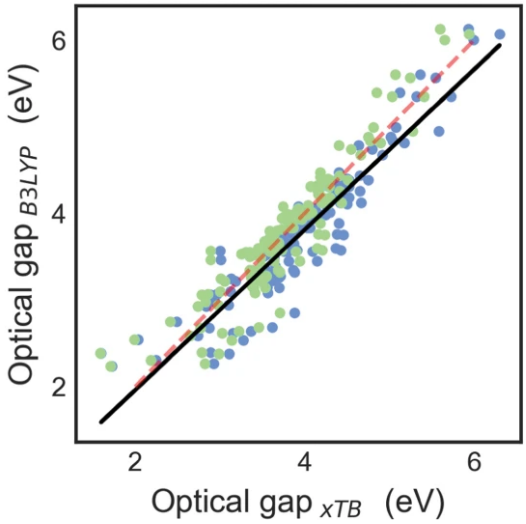


FIGURE 4.1: Linear calibration of S1 calculated by xTB-sTDA vs. TD-DFT (B3LYP). Blue points show original data while green points show calibrated data. Black line is linear fit of original data while red line is $x = y$ line. From "Mapping the opto-electronic property space of small aromatic molecules" by L. Wilbraham et al., 2020, *Communications Chemistry* volume 3, Article number: 14.[112]

Thus, instead of using a linear calibration model, this work proposes training a machine learning model for calibration of excitation energies output by xTB-sTDA. Theoretically, an ML model should detect higher order patterns than a linear calibration model would. There is also precedent for calibration ML models in literature, as presented in Section 2.2.2. The following section presents the methodology used to generate a calibration ML model. First, we compare several ML models to choose the best model architecture. We then generate a training set and discuss the workflow used to train the ML model.

4.2 Methodology

4.2.1 Comparing ML models

The 3 ML models considered were DeepChem's[146] GCN[147] (DC GCN), DeepChem's MPNN[148] (DC MPNN), and Chemprop's MPNN[123] (CP MPNN). The default, out-of-the-box settings for each ML model were used, as described in Appendix Section A.3. The input for the ML models was a CSV file with 3 columns: the SMILES representation of the molecule, the S1 error between xTB-sTDA and TD-DFT, and the T1 error. The goal of each ML model was to accurately predict the error between xTB-sTDA and TD-DFT for a given molecule.

A test set was required to test different ML models. The VERDE materials database (VerdeDB) was used for this purposes, as it is a carefully curated database for excited-state properties of organic molecules.[93] VerdeDB contains three classes of molecules: porphyrins, quinones, and dibenzoperylenes, which are commonly used in applications in renewable energy and green chemistry.[93] Of the 1500 molecules, around 1000 had both S1 and T1 energies available, so these were used as the test set. xTB-sTDA was run on all 1k molecules (details of the workflow for running xTB-sTDA are presented later in Section 4.2.3). The S1 and T1 errors between xTB-sTDA and TD-DFT were then calculated and tabulated.

Instead of predicting both S1 and T1 error simultaneously, two separate single-task models were generated. 10-fold cross-validation was conducted by splitting the VerdeDB data 80%/10%/10% into train/validation/test sets. For each fold, the trained ML model was used to predict error values of the test set. Then, each molecule's predicted error was added to the xTB-sTDA output to give a calibrated energy, called the xTB-ML value. The xTB-ML values were compared to the TD-DFT reference results by calculating an R2 score.

Figure 4.2 shows the results of comparison for T1 and S1 energies. As seen, all ML models vastly outperform the linear calibration method. Between the ML models, CP MPNN performs the best for both T1 and S1, with an average R2 of 0.89 for T1 calibration and 0.77 for S1 calibration. Note that the large variability in R2 can be explained by the presence of outliers in the test set - since the test set was only composed of 100 molecules (10% of 1k), a few outliers can vastly impact performance.

Figure 4.3 shows plots of original vs. CP MPNN-calibrated xTB data for (a) T1 and (b) S1 energies, with test data from all 10 folds compiled and with outliers removed. The accuracy of the calibrated data is high, with an R2 of 0.964 and MAE of 0.098 for T1 and R2 of 0.851 and MAE of 0.130 for S1.

From this analysis, it is evident that CP MPNN performs well in calibrating xTB results, even with its default settings. To see if the performance could be boosted further, various improvements were attempted. These included increasing the number of epochs to 100, conducting hyperparameter optimization, adding RDKit-calculated features as input (in addition to the NN-calculated features), and conducting multi-task training. The results from these improvements are shown in Figure 4.4.

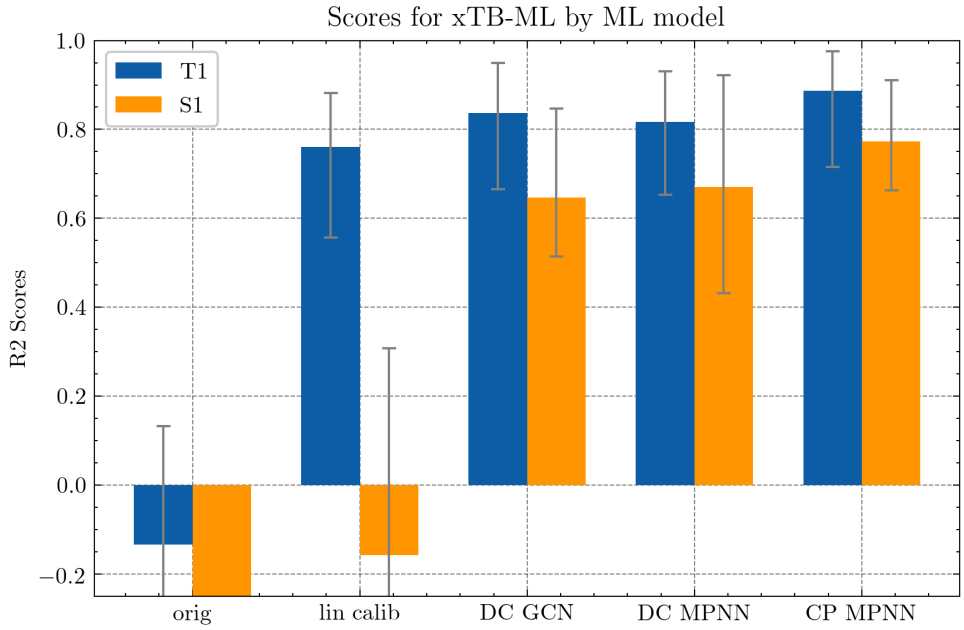


FIGURE 4.2: Comparison of various ML models in accurately calibrating xTB against TD-DFT, quantified by R2 score. 'orig' = original xTB data with no calibration, 'lin calib' = linear regression calibration of xTB data. All others are ML models as presented above. Blue bars are xTB-ML T1 energies while orange bars are xTB-ML S1 energies. R2 for original S1 data is -1.84 $\pm$ 0.65, the plot was truncated for clarity.

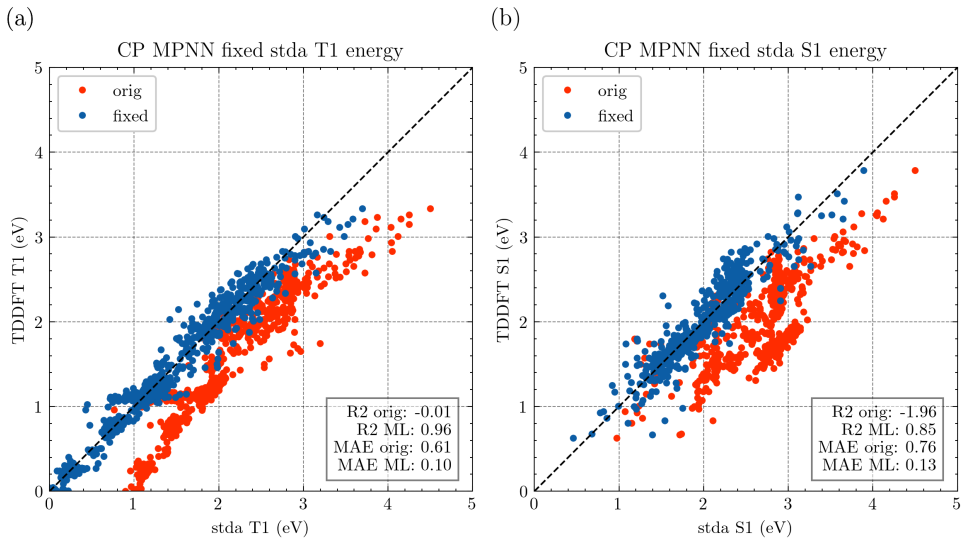


FIGURE 4.3: Plot of original xTB data ('orig', red) and CP MPNN ML-calibrated xTB data ('fixed', blue) against reference TD-DFT data generated with Gaussian, for (a) T1 energies and (b) S1 energies. Datapoints are all test data compiled across 10 non-overlapping folds in cross-validation.

As seen, there are only small differences in performance between the default settings and any potential improvements to the ML settings. For T1, hyperparameter optimization provides minimal improvement, while including additional features or adding multitasking reduces accuracy. For S1, hyperparameter optimization marginally improves performance, and adding multitasking also seems to improve

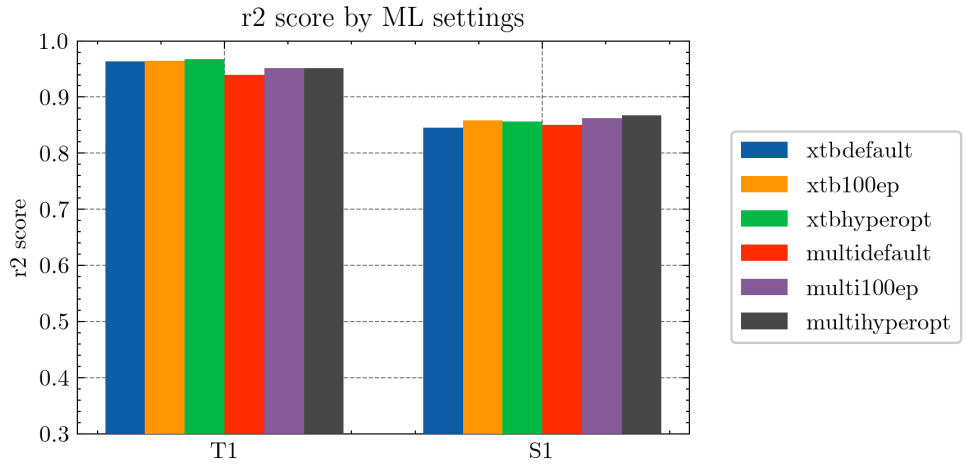


FIGURE 4.4: R2 scores of xTB-ML vs. TD-DFT for various improvements attempted to CP MPNN. Bars labeled 'xtb' are single-task and require 2 different models to predict S1 and T1, while bars labeled 'multi' are multitask and only 1 model predicts both S1 and T1. 'default' bars use the out-of-the-box hyperparameter settings with no additional features. '100ep' bars use 100 epochs instead of the usual 30. 'hyperopt' bars use hyperparameter optimization. 'rdkit' bars include RDKit-calculated features as additional inputs.

performance. There is thus a tradeoff in using multitasking as it could reduce accuracy for T1 predictions but improve accuracy for S1, while also reducing overall computation time. Because of the time savings of the multi-task model, this was used for ML for the following sections. Hyperparameter optimization was ruled out due to the operation being too expensive - optimization had to occur for each fold independently and took several times longer than the actual ML run, while only providing marginal improvements.

We have therefore chosen the multi-task CP-MPNN as our ML model architecture. We can now develop a larger-scale ML calibration model. While the VerdeDB database above was useful for comparison, it is small and not very diverse. Thus, we need to generate a larger training set of molecules for which excited states are relevant, as discussed in the following section.

4.2.2 Dataset descriptions

In order for the resulting ML model to be accurate and widely applicable, we need to generate a large, versatile training dataset. Ideally, this would include molecules similar to those of interest. Excited state energies such as singlet and triplet states are currently of interest in triplet-triplet annihilation (TTA) and singlet fission (SF) materials, so molecules involved in these processes would be best to include in the training set. TTA/SF occurs at a wide range of energies, so it is important to not be restrictive to certain energies in the training dataset.

For the test dataset, while it is always possible to choose a subset of the training dataset, it would be better to additionally have a blind test dataset to evaluate

the generalizability of the ML model. Details regarding training and test dataset generation are provided below.

SCOP-PCQC: Literature scraping of relevant molecules

As mentioned above, it would be ideal to include molecules involved in TTA/SF in the training dataset. However, there is no existing database of such molecules, so independent generation of such a database was necessary. The key characteristics of this database are molecule name/information, as well as excited state energy data.

To get the names of molecules involved in TTA/SF, we created a literature scraping workflow. We used the SCOPUS API[149] to obtain abstracts of articles tagged with TTA/SF keywords. Then, we used ChemDataExtractor[150] to extract molecule names from the abstracts. We then used the PubChem API[151] to convert molecule names into PubChem CIDs. Finally, the PubChem API was used again to conduct a 2D Tanimoto-coefficient based similarity search among PubChem molecules to expand the molecular space of interest. Overall, this process allowed us to get molecular information for all relevant molecules.

To obtain excited state energy data, we cross-referenced all of the PubChem CIDs against PubChemQC (PCQC),[65] a database of various quantum chemistry properties of 3.5 million molecules, including the S1 energy. The T1 energy was not included in PCQC, so it was independently generated with TD-DFT, by including the `triplets` keyword in the Gaussian file provided by PCQC and running it on our own cluster.

The final count of this portion of the training set (named SCOP-PCQC) was approximately 10k molecules. Figure 4.5 shows the workflow used to generate the SCOP-PCQC dataset.

As described above, SCOP-PCQC is a subset of the PubChemQC dataset. PubChemQC is itself a subset of all PubChem molecules, with some restrictions (only certain atoms allowed, no SMILES containing periods, ignoring isotopes, neutral charges) as detailed in Nakata and Shimazaki.[65] Figure 4.6 shows plots of some properties of PubChemQC molecules in comparison to all PubChem molecules. Figure 4.7 shows plots of properties of SCOP-PCQC molecules in comparison to PubChemQC molecules.

As seen in Figure 4.6, due to the restrictions in SMILES included in PubChemQC, the molecules are relatively simple, with low molecular weight (MW), complexity, heavy atom count, and number of rotatable bonds compared to all molecules in PubChem. There is a wide variability in S1 energies in PubChemQC, in a bell curve shape centered around 5 eV.

From Figure 4.7, we can see SCOP-PCQC is a representative subset of all of PubChemQC. As seen in Figure 4.7(a), SCOP-PCQC seems to include the "core" molecules, leaving out molecules with high complexity but low MW or vice versa. SCOP-PCQC also covers all the options in Figure 4.7(b). Finally, SCOP-PCQC has a similar distribution of S1 energy, as seen in Figure 4.7(c).

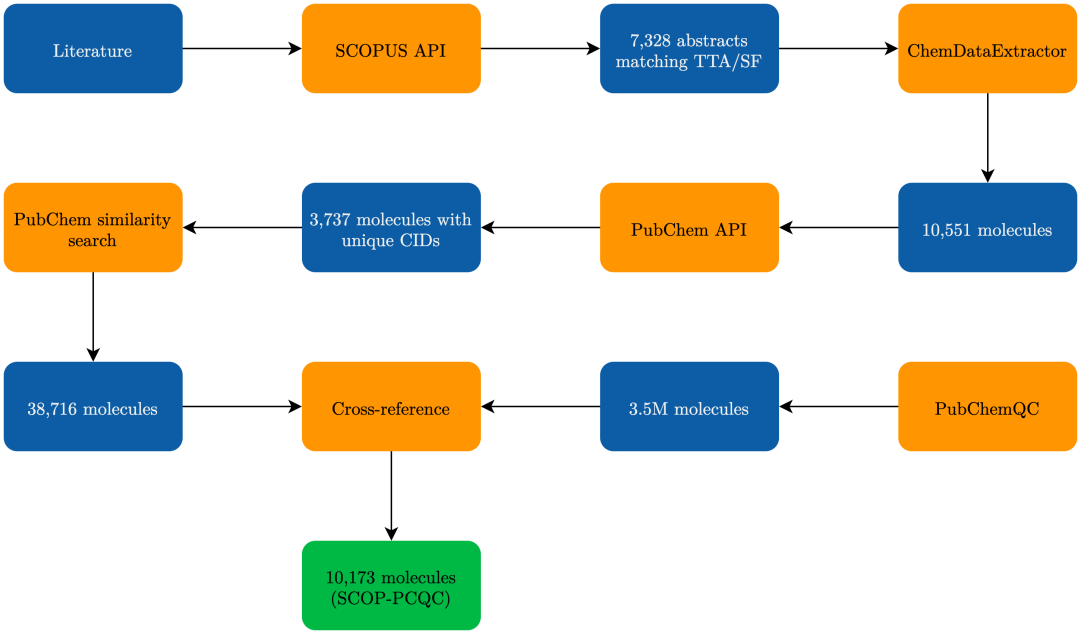


FIGURE 4.5: Workflow used to generate the SCOP-PCQC dataset, starting with literature scraping for relevant molecules and ending with molecular information and excited state energy data for 10k molecules. Blue boxes indicate data while orange boxes indicate methodology. Green box indicates final data in SCOP-PCQC dataset.

It is useful here to conduct an analysis of the specific molecular substructures of the SCOP-PCQC dataset, to understand the components and diversity of this training dataset. This is done in two ways - first by searching for specific substructures in SCOP-PCQC, and second by mapping and clustering of molecules in chemical space.

First, specific substructures known to be relevant to TTA-SF processes were searched for in the dataset. Pyrene showed up as a substructure in 41 molecules, perylene in 4, anthracene in 107, and naphthacene in 4. For a more comprehensive analysis, the 143 molecules used for calibration in Wilbraham et al.[112] (chosen as a representative sample of small aromatic molecules, discussed further in Section 4.2.2) were selected as a test set of substructures. Of the 143 substructures, 68 had matching molecules in SCOP-PCQC. Among the 68 substructures, an average of 165 matching molecules were found in SCOP-PCQC. A boxplot of number of matching molecules for the 68 substructures is shown in Figure 4.8. Because the mean (165) is higher than the median (15), there are a few substructures with many matches in SCOP-PCQC, creating a right-skewed distribution.

For further substructure analysis, it is possible to extract the scaffold (core structure) of each molecule in SCOP-PCQC and count the number of occurrences of each scaffold in the dataset. This helps understand the diversity of chemical substructures in SCOP-PCQC. RDKit's Murcko scaffold implementation was used used for this analysis. Figure 4.9 shows the 98 most common substructures in SCOP-PCQC, each occurring in at least 10 molecules. We can again see the pyrene and anthracene substructures here, as well as a variety of other diverse substructures, including many

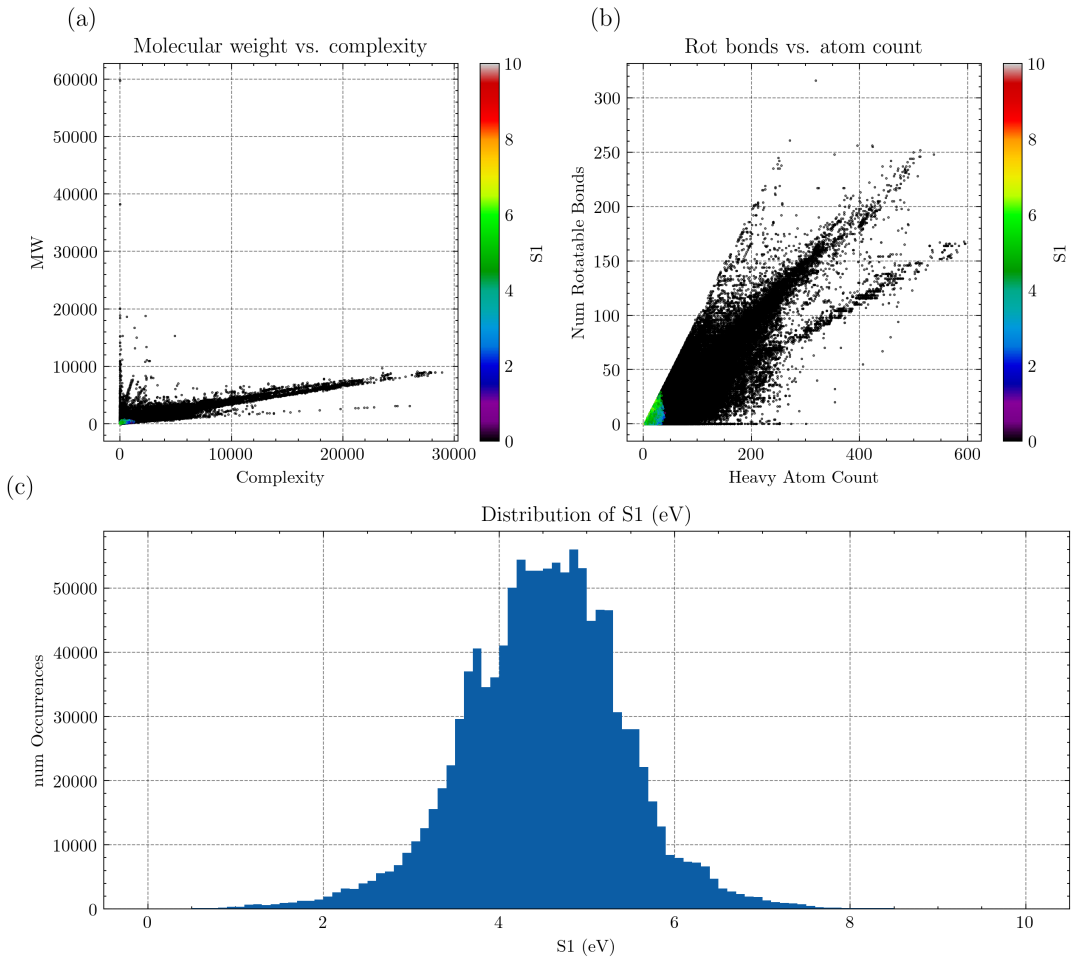


FIGURE 4.6: Plots of properties of molecules in the PubChemQC database. (a) Shows MW of molecules vs. complexity. Complexity is evaluated both by elements contained and structural features including symmetry.[152] (b) Shows number of rotatable bonds vs. heavy atom (non-hydrogen) count. For both (a) and (b), black dots represent molecules in PubChem, while colored dots represent molecules in PubChemQC – color is based on the S1 energy value. (c) Shows a histogram of all S1 energies from the PubChemQC database.

aromatics. We can thus be confident in the accuracy of our ML model for similar molecules.

The second way of analyzing the molecular composition of the SCOP-PCQC dataset is through chemical space mapping. A UMAP embedding of 350k molecules subsampled from the PCQC dataset is first generated, and the 10k molecules in SCOP-PCQC are embedded into this global chemical space, as shown in Figure 4.10. This is to show the distribution of molecules in global chemical space, and to get a sense of overall coverage.

Next, the SCOP-PCQC molecules are embedded with t-SNE, and HDBSCAN is used for clustering, as seen in Figure 4.11. Clustering allows further analysis of molecular substructures, as molecules with similar skeletons would be neighbors in chemical space. For substructure analysis, the molecules are split into 1000 clusters, and the maximum common substructure (MCS) of each cluster is calculated. If the

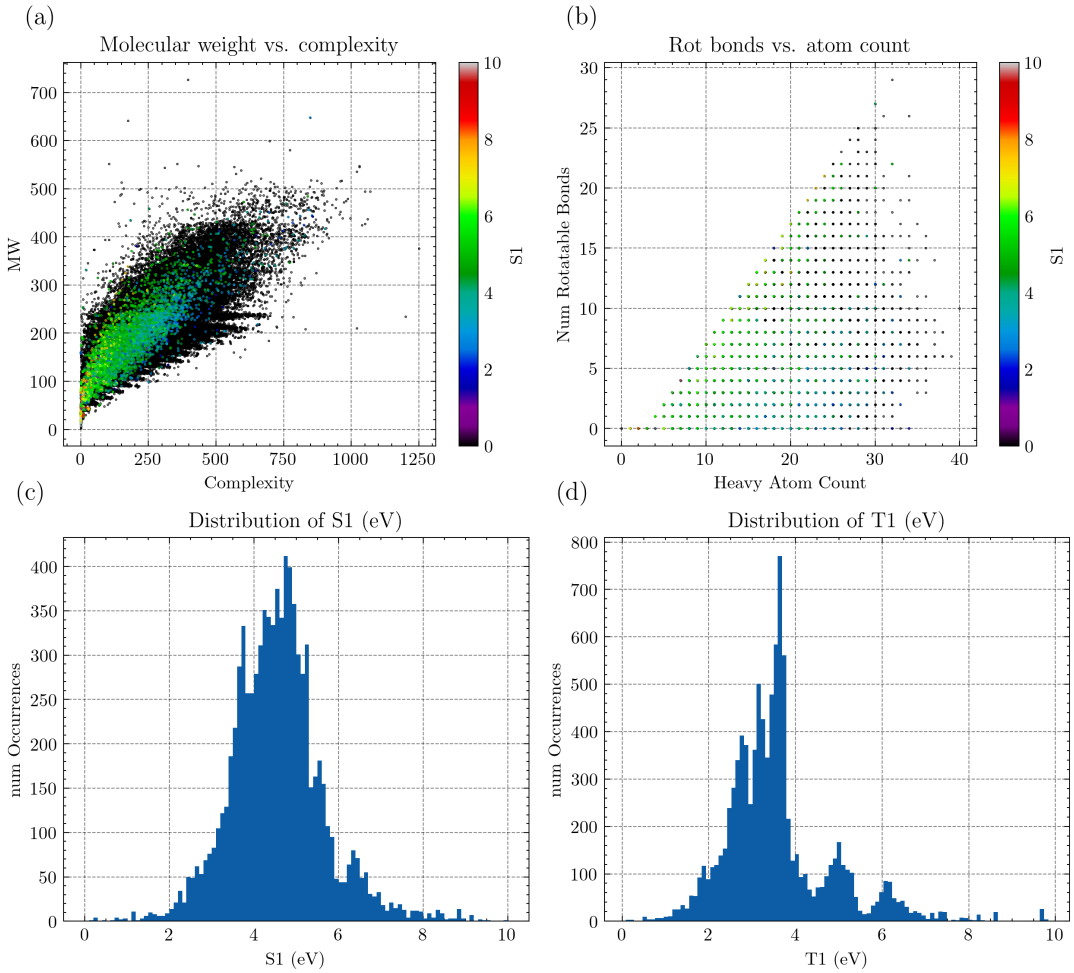


FIGURE 4.7: Plots of properties of molecules in the SCOP-PCQC dataset. *x*- and *y*- axes of plots are the same as in Figure 4.6. For (a) and (b), black dots represent molecules in PubChemQC, while colored dots represent molecules in SCOP-PCQC – color is based on the S1 energy value. (c) Shows a histogram of S1 energies of SCOP-PCQC molecules from the PubChemQC database. (d) Shows a histogram of T1 energies calculated independently with TD-DFT in Gaussian.

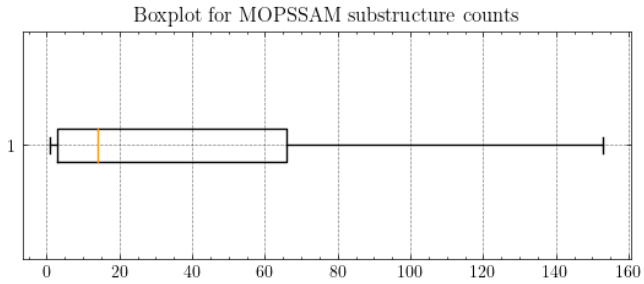


FIGURE 4.8: Boxplot of substructure matching counts of 68 substructures (selected from Wilbraham et al.'s[112] calibration set) in SCOP-PCQC. Shows the inclusion of important aromatic substructures in the SCOP-PCQC training dataset.

MCS has >10 atoms, it is shown in Figure 4.12.

As seen, the SCOP-PCQC dataset includes a diversity of molecules. From Figure 4.10, it is evident the dataset broadly covers the global chemical space. There

are some areas of greater concentration, for example the left center and right lower regions. This indicates more molecules in these regions are of interest in TTA/SF, perhaps because existing TTA/SF molecules are derived from the same classes of molecules. To better understand these clusters, Figure 4.12 shows the specific MCS of various clusters, labeled by their cluster number. This is useful to understand which molecular substructures correspond to specific regions of chemical space.

However, there are a few limitations of the SCOP-PCQC dataset. First, although the dataset broadly covers the chemical space, there are a few areas with gaps, and a few areas with higher concentration of molecules. This could be because the dataset was generated from molecules relevant to TTA/SF, so there is a potential for it to be homogeneous. Second, while it contains molecules with a wide variety of S1 energies, there are only a few with low S1 energy (<2 eV). These low-S1 molecules would be more relevant to solar applications as typical solar cells have a bandgap of around 1.1-1.5 eV. Thus, we need to supplement SCOP-PCQC.

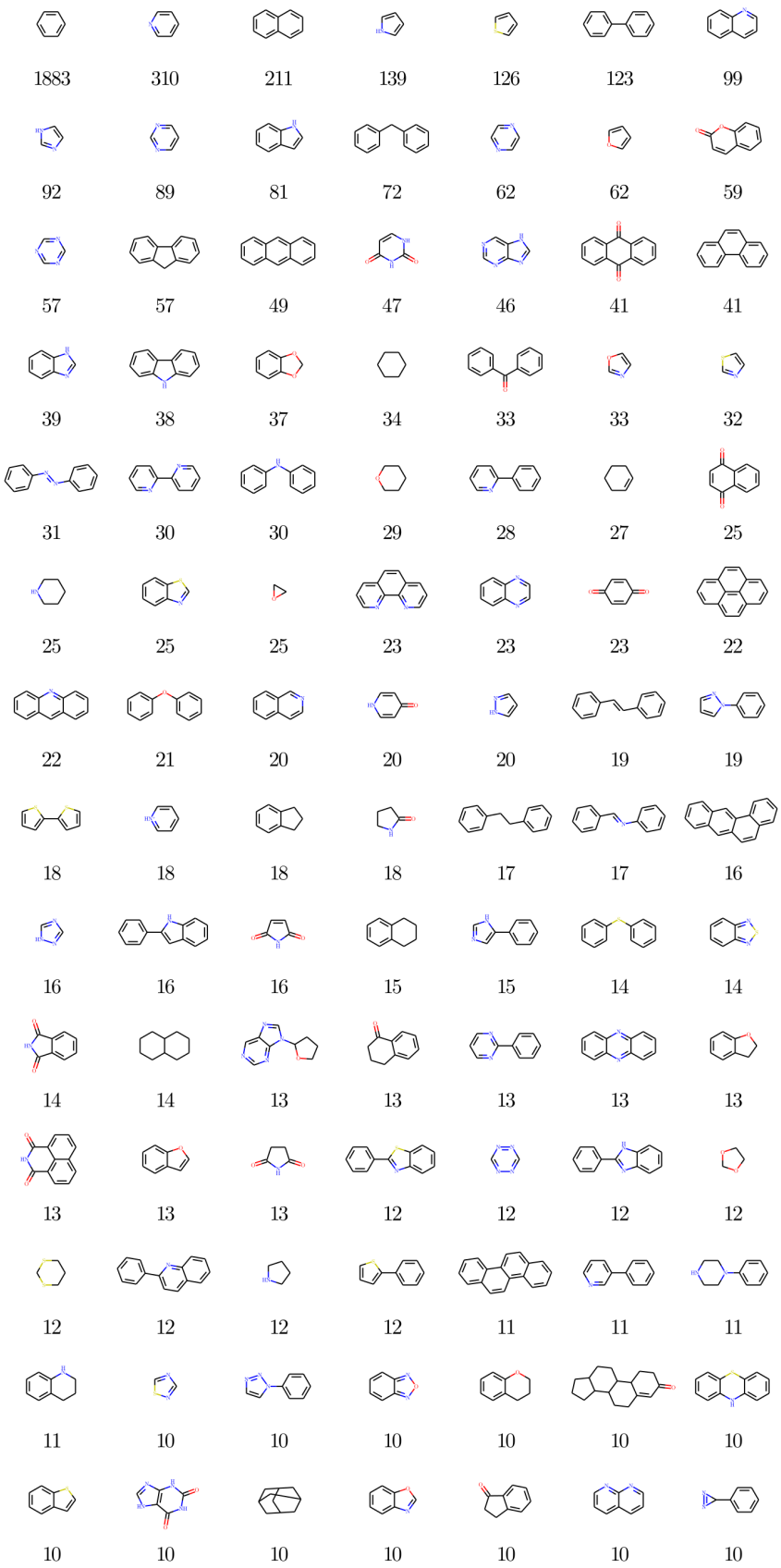


FIGURE 4.9: Most common scaffolds in the SCOP-PCQC dataset, with the number of occurrences displayed below each scaffold. Scaffolds generated with RDKit's Murcko scaffold implementation.

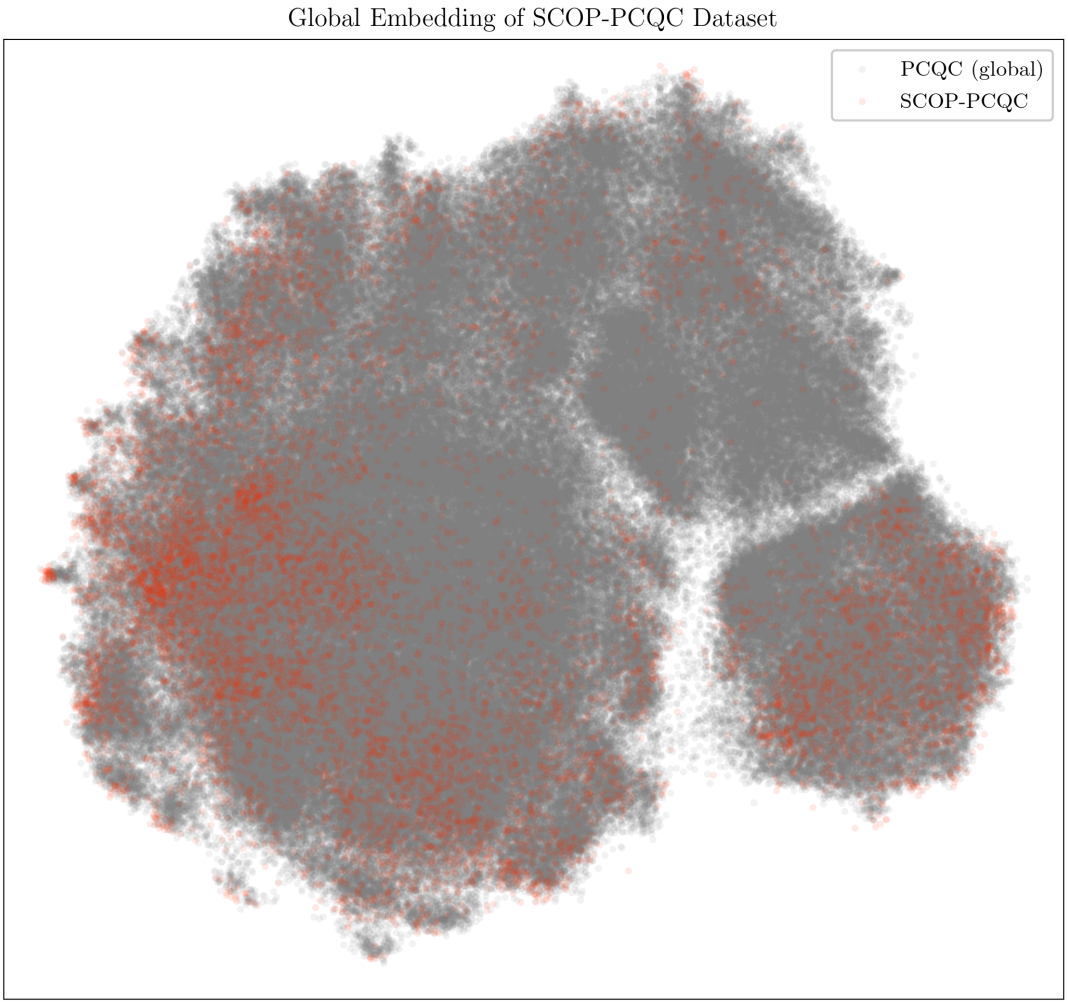


FIGURE 4.10: Embedding of SCOP-PCQC data (red) in the global PCQC chemical space (grey). Shows broad coverage of the global chemical space with some areas with greater concentration of molecules. A UMAP model created on 350k molecules of PCQC was used to predict the locations of the 10k SCOP-PCQC molecules. UMAP was chosen for this embedding due to its speed - tSNE would be too slow for this many molecules.

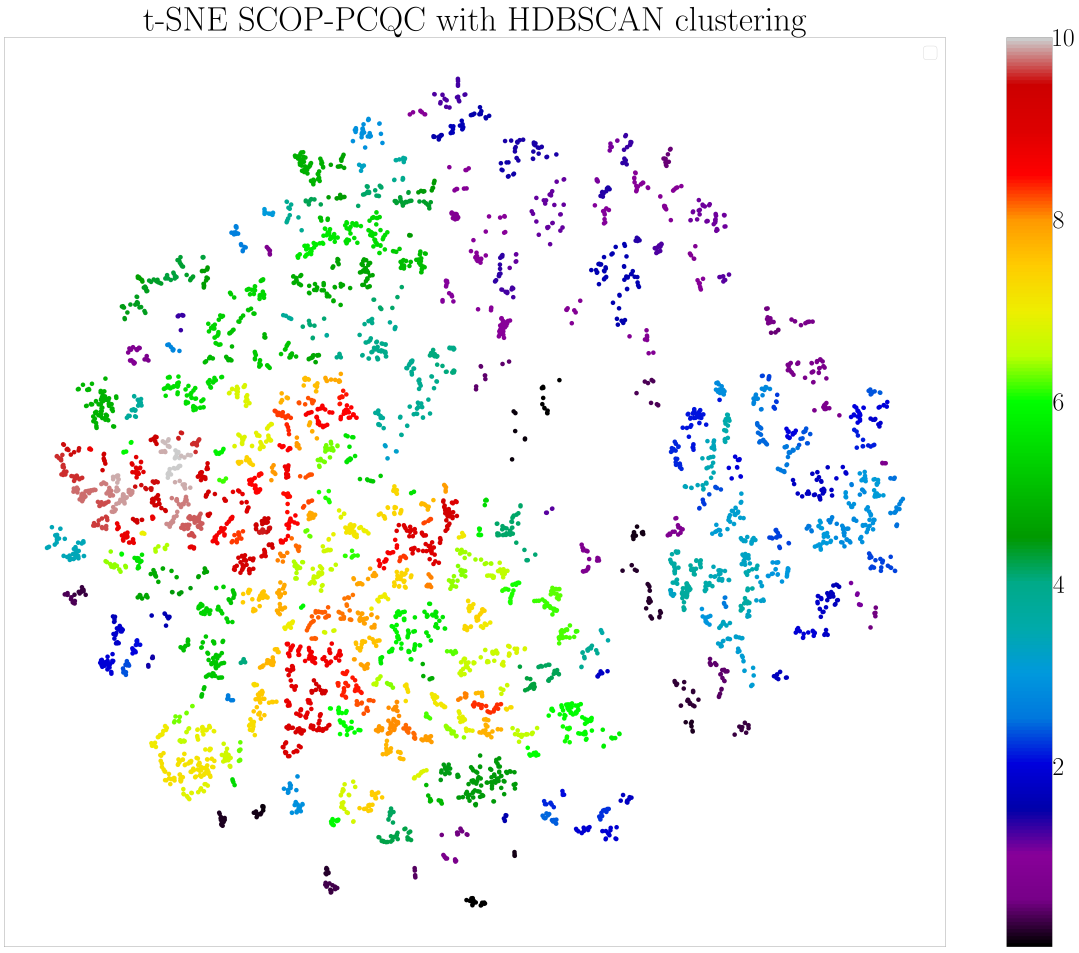


FIGURE 4.11: t-SNE embedding of 10k SCOP-PCQC molecules with HDBSCAN used for clustering. Partial global embedding was done with 10k molecules subsampled from PCQC for reference (not shown). Note that the overall structure of the chemical space resembles the UMAP embedding. HDBSCAN generates a soft cluster with float values ranging from 0 to 10. Of the 10k molecules in the dataset, approximately 7.5k were able to be clustered and are shown in the Figure.

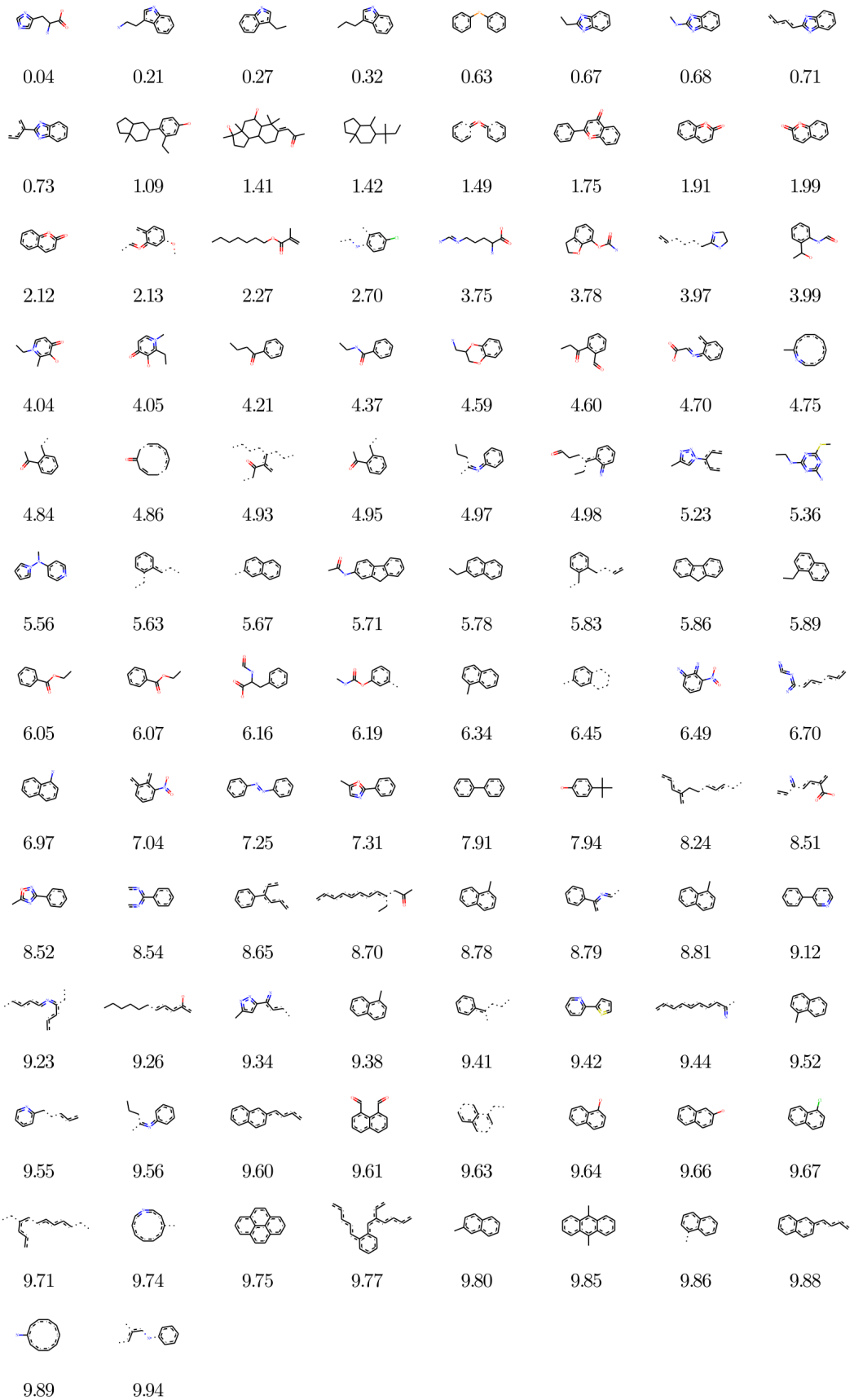


FIGURE 4.12: Maximum common substructures (MCS) of the 98 (/1000) clusters that had MCS with more than 10 atoms. Number below each MCS corresponds to the cluster number, ranging from 0 to 10, corresponding to the colorbar in Figure 4.11

SCOP-PCQC Expansions

The first supplement to SCOP-PCQC is to add low excited state energy molecules from PCQC. This is done by splitting the molecules in PCQC into bins of width 0.1 eV, and sampling 100 molecules from each bin between 0 and 2.5. Figure 4.13 shows the results of this expansion.

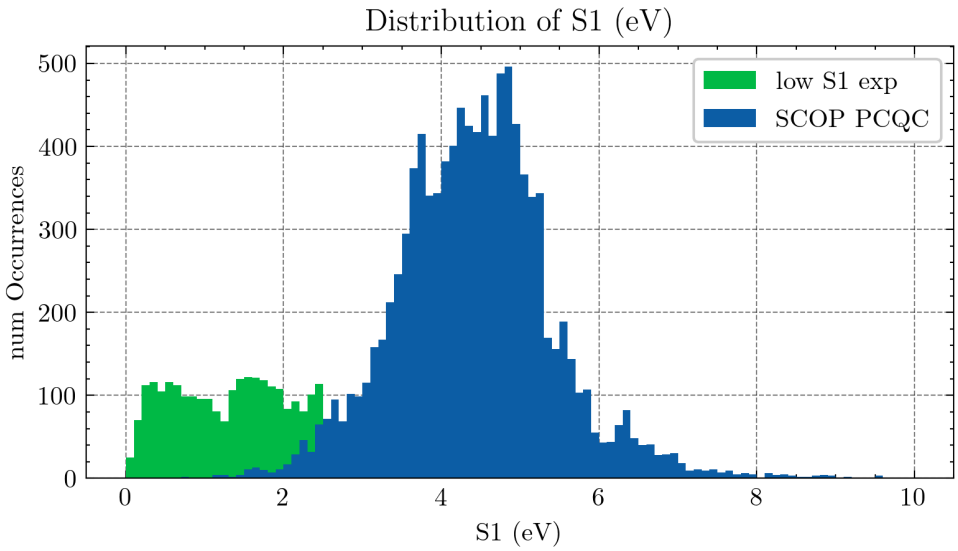


FIGURE 4.13: Histogram showing the expansion of SCOP-PCQC to include low-S1 molecules. Blue bars indicate original SCOP-PCQC data while green bars indicate low-S1 expansion data.

While this addition was motivated by visually inspecting the S1 distribution, this may not be the most effective methodology. For example, some molecules may be similar to existing molecules in the training set, reducing their utility, or the additions could be too clustered in chemical space to give a useful addition. To more intelligently expand the SCOP-PCQC dataset, we turn to our analysis in Chapter 3, with active learning for dataset generation. In this section, we supplement the initial SCOP-PCQC dataset based on the active learning workflow, i.e. generating an ML model from SCOP-PCQC and using the model to measure epistemic uncertainty on the remaining PCQC molecules. In this way, we get a sense of which molecules the ML model has difficulty predicting, and can add these molecules to the training set.

As discussed in Chapter 3, the AL workflow was run separately for S1 (Section 3.3.1) and T1 (Section 3.3.3) predictions, giving two different ML models. The molecules in the first AL cycle for both S1 (64k molecules) and T1 (107k molecules) were also used as supplements to SCOP-PCQC.

QM-symex-10k

To ensure broader applicability of our model beyond the PCQC dataset, it was important to include another source of data in our training set. As seen in Section 2.1.1, unfortunately there are only a few excited state databases available, especially

for triplet energies. While it is possible to independently generate triplet data, as we did with SCOP-PCQC, to save computation time it would be best to choose a database with triplet energies calculated. Of the triplet energy databases available, QM-symex is the largest and most versatile. To balance the SCOP-PCQC dataset, 10k molecules were randomly sampled from QM-symex to form the QM-symex-10k dataset. Figure 4.14 shows plots of some properties of the QM-symex-10k dataset, including molecular weight, complexity, number of rotatable bonds, heavy atom count, and S1/T1 energies.

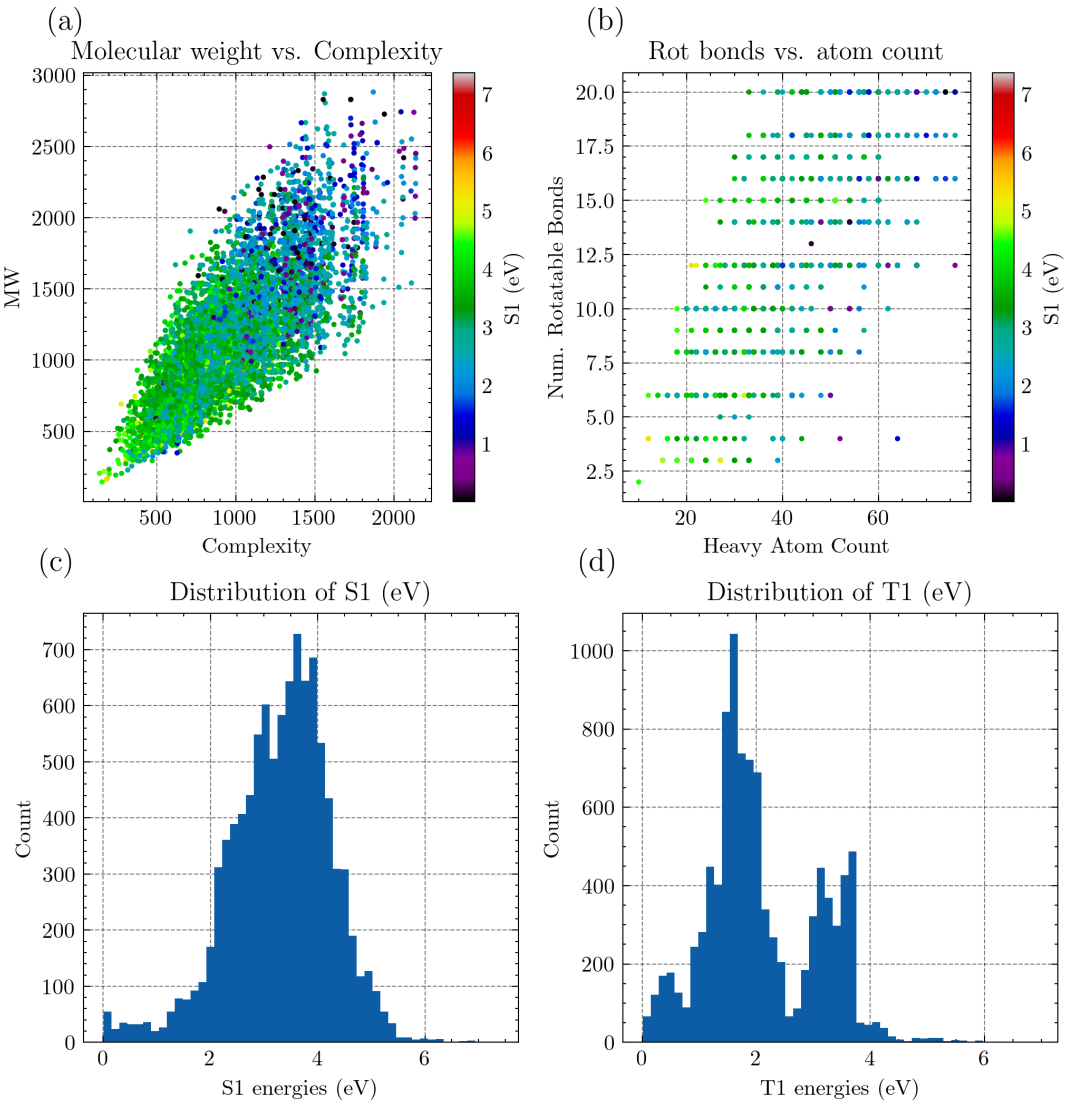


FIGURE 4.14: Plots of properties of molecules in the QM-symex-10k dataset. (a) Scatter plot of molecular weight vs. complexity, colored by S1 energy. (b) Plot of number of rotatable bonds vs. heavy atom count (non-H), also colored by S1 energy. Histograms of (c) S1 and (d) T1 energies.

As seen in Figure 4.14, QM-symex-10k includes more molecules with low excitation energies, especially for T1. Quantitatively, QM-symex-10k has 776 molecules with S1 < 2 eV and 5877 molecules with T1 < 2 eV. In addition, QM-symex-10k has larger and more complex molecules – SCOP-PCQC has a maximum complexity of

around 750 and a maximum MW of around 400, but QM-symex-10k molecules reach complexities of 2000 and MWs of 2500, as well as heavy atom counts of up to 80. Finally, QM-symex-10k helps increase the diversity of the dataset as it does not explicitly include molecules relevant for TTA/SF.

Now that we have our training datasets, we have to generate a test dataset to evaluate the accuracy and applicability of our ML model.

Blind test datasets

As discussed in Section 4.1.1, one of the few papers in xTB calibration against TD-DFT data was by Wilbraham et al.,[112] in which they calibrated IP, EA, and S1 data output by xTB against TD-DFT data with a linear model. It therefore makes sense to use this calibration dataset as a test set for our study, to see if our ML methods can improve the calibration. This dataset is called "MOPSSAM" as an acronym of their study ("Mapping the optoelectronic property space of small aromatic molecules"[112]) and contains xTB-sTDA calculated, linearly calibrated, and TD-DFT calculated S1 energy for 143 molecules. Since we are interested in triplet energies, T1 energies were independently calculated for the 143 molecules. To verify the workflow of TDDFT calculations in this study was equivalent to that used in Wilbraham et al., S1 energies were first calculated independently and compared to the previous work. As shown in Appendix Figure A.2, there is very little deviation in our results. When calculating T1 energies, the only difference in workflow is adding the `triplets` keyword in the Gaussian input. Therefore, we can be certain that the generated T1 data is valid. This preliminary test set is called MOPSSAM 143.

To expand the test dataset, 1k additional molecules were randomly chosen from the 250k molecules considered in their study, and S1/T1 energies were independently calculated for these molecules with TDDFT. This expanded dataset is called MOPSSAM 1000.

Finally, to test the broader applicability of the model, molecular excited state information was taken from Fallon et al.'s paper on designing singlet fission materials based on indolonaphthyridine thiophene (INDT) derivatives.[52] As a singlet fission study, it featured S1 and T1 calculations with TD-DFT using B3LYP/6-31++G** and the Tamm-Dancoff approximation on almost 10k molecules.[52] Because the molecules included in the INDT dataset are substantially different from MOPSSAM, this should be a useful test set to determine the generalizability of the ML model.

With these training and test datasets, an ML model can be generated for calibration of xTB results, as detailed in the following section.

4.2.3 xTB-ML calibration workflow

An overview of the xTB-ML calibration workflow is presented in Figure 4.15. Since the training datasets already have TD-DFT calculated S1 and T1 data, we can directly extract these, but we need to generate xTB-sTDA data. Starting with the SMILES

string, we generate an initial 3D molecular structure with OpenBabel's `gen3d` function.[124] This function has 4 parts: (i) preliminary 3D structure generation with OBBuilder, (ii) 250 steps of steepest descent geometry optimization with MMFF94, (iii) 200 iterations of a conformer search, with 25 steps of steepest descent optimization for each conformer, and (iv) 250 steps of conjugate gradient geometry optimization on the lowest energy conformer.[153,154] We then optimize the initial 3D structure with xTB geometry optimization using GFN2-xTB and a `tight` threshold. The resulting final 3D ground-state structure is then run through xTB-sTDA to get excited state (S1, T1) data. Then, the error is calculated between TD-DFT and xTB-sTDA. Finally, the error and SMILES string for each molecule are passed to the ML model as input.

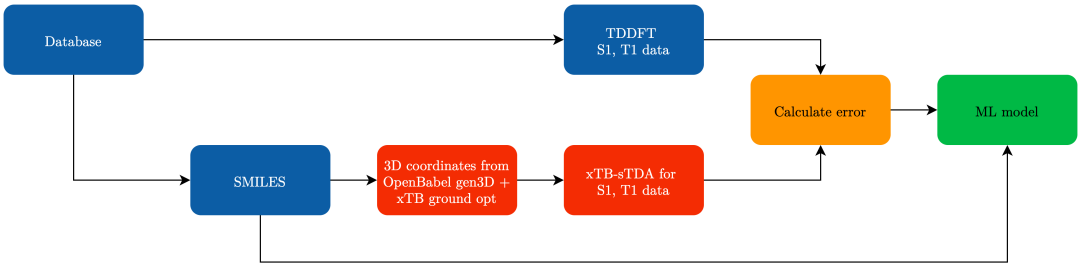


FIGURE 4.15: Workflow for xTB-ML calibration. Blue boxes represent data, red boxes represent intensive calculations, yellow boxes represent quick calculations, and green box represents final result. Starting with the training datasets, the TD-DFT and SMILES data are directly extracted. The SMILES strings are converted to 3D molecular structures with OpenBabel and xTB, and then excited state calculations are conducted with xTB-ML. Then the error is calculated between xTB-sTDA and TD-DFT and fed as input to the ML model, along with the SMILES string.

Note here that even though a 3D structure must be generated in order to get xTB-sTDA data, it is not used as input in the ML model. This is due to the characteristics of ML model used (Chemprop's MPNN) which simply takes SMILES strings as input. An avenue for future work could be to use the full 3D molecular structure as input to the ML model.

While the xTB portion of this workflow is standardized, the TD-DFT portion may not be, as discussed in the following subsection.

Comparison of TD-DFT settings

The TD-DFT data is sampled directly from existing databases, but the steps used to generate this data can differ. These differences include: (a) different ways of generating initial 3D coordinates, (b) different DFT settings used to optimize the ground state geometry, and (c) different TD-DFT settings (such as functionals, basis sets, optimization tightness, and solvation models) used to calculate excited state energy. Table 4.1 compares these settings for the 3 databases considered in this work (PubChemQC, QM-symex, and MOPSSAM).

As seen in Table 4.1, the three databases use different techniques in generating excited state data. They all use different methods for generating 3D coordinates and conducting conformer analysis. There are also similarities, however, as they all use

TABLE 4.1: Comparison of settings for the TD-DFT workflow in the 3 databases considered in this study. The 3 databases use different 3D coordinate generation techniques, but all use the B3LYP functional for (TD-)DFT, despite using different basis sets. MOPSSAM is the only study that used a solvation model.

	PubChemQC	**QM-symex**	**MOPSSAM**
3D coord gen	OpenBabel PM3 STO-6G	Corina PM7	RDKit EKTDG MMFF94
DFT	B3LYP 6-31G(d)	B3LYP 6-31G(2df,p)	B3LYP aug-cc-pVTZ
TD-DFT	B3LYP 6-31+G(d)	B3LYP 6-31G	B3LYP aug-cc-pVTZ
Solvation	None	None	Benzene COSMO

the B3LYP functional for (TD-)DFT, despite using different basis sets. MOPSSAM is the only dataset that uses a solvation model.

While the goal in this study is to combine the two training sets (SCOP-PCQC and QM-symex-10k) into one overarching ML model, it is unclear whether the differing settings between the two datasets will impact results. This is why we use MOPSSAM as a blind test set to determine the applicability of the generated ML model.

Based on the workflow and considerations outlined above, input data can be generated for the ML model. The following section provides details of the ML models generated and results for calibration accuracy.

4.3 Results

First, ML models are trained separately on the SCOP-PCQC and QM-symex-10k datasets to determine the accuracy of calibration independently on these datasets. 10-fold cross-validation results are presented in the first two subsections. As a consequence of cross-validation, the test sets are subsets of the entire dataset. To test the broader applicability of our methodology, an overarching ML model is trained with all 20k molecules and blind tested on MOPSSAM data, as presented in the third subsection. Finally, expanded ML models are trained and blind tested on MOPSSAM data, also presented in the third subsection.

4.3.1 SCOP-PCQC

Figure 4.16 shows the results of 10-fold cross-validation calibration on the SCOP-PCQC dataset, for both S1 and T1 energies. While the original data has low accuracy when compared to TD-DFT results, the linear calibration improves the accuracy slightly. However, there is not a clear linear shift due to some groups of molecules located farther from the line of best fit. This creates large errors for some molecules – this is most clearly seen for xTB T1 energies between 5-6 eV and between 6-7 eV for TDDFT.

Adding ML boosts the accuracy further. This is likely because ML allows for higher-order pattern detection, allowing groups of molecules to shift locally instead of having to follow a global calibration rule. The MAE for ML-calibrated xTB-sTDA for both S1 and T1 is ∼0.20 eV. The RMSE (not listed) is ∼0.37 eV for both.

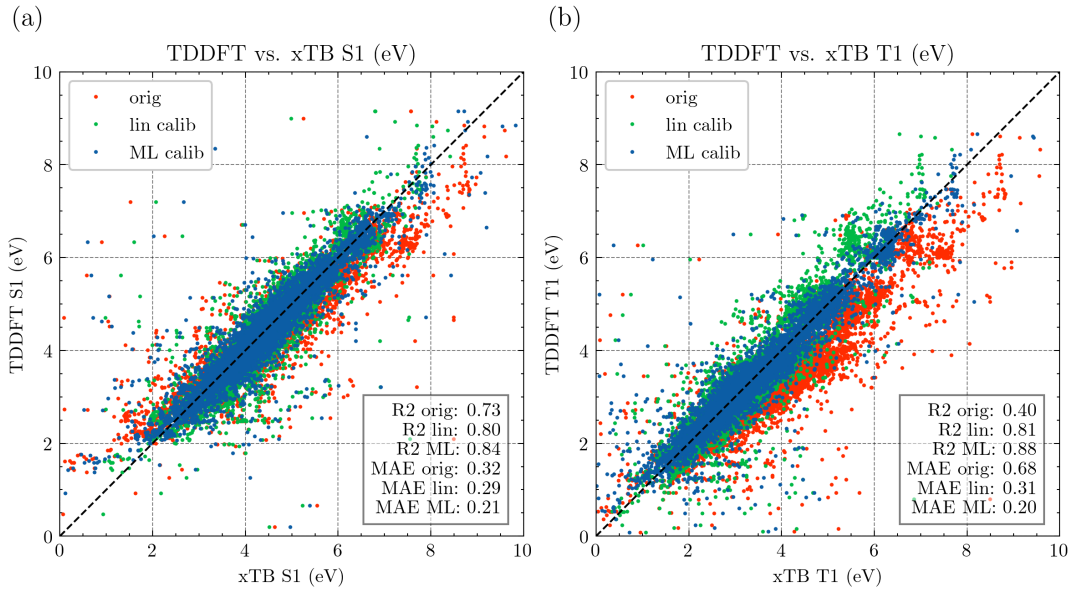


FIGURE 4.16: Plots of xTB calibration of the SCOP-PCQC dataset for (a) S1 and (b) T1 energies. Red dots are original data with no calibration, green dots are linearly calibrated data, and blue dots are calibrated with ML. 10-fold cross-validation was conducted, meaning all data points shown are test points predicted by an ML model trained on the other 90% of data. Inlaid box shows quantitative measurements of accuracy for original, linearly calibrated, and ML calibrated data. (Best R2 is 1 while best MAE is 0.)

4.3.2 QM-symex-10k

Conducting the same analysis on QM-symex-10k produces similar results. Figure 4.17 shows the original data as well as the results of 10-fold cross-validation on linear- and ML-calibrated xTB-sTDA. Again, the linear calibration has some significant limitations, with some datapoints not being able to shifted enough (clearly seen for low TDDFT S1 energy) or shifted too much (clearly seen for xTB S1 energies around 4eV and TDDFT S1 energies above 4eV). The T1 plot similarly shows some groups of molecules that are clearly shifted away from the line of best fit.

The ML calibration boosts accuracy for both S1 and T1, likely for the reasons discussed above. The MAE for ML-calibrated xTB-sTDA for both S1 and T1 is ∼0.20 eV, while the RMSE (not listed) is ∼0.30 eV.

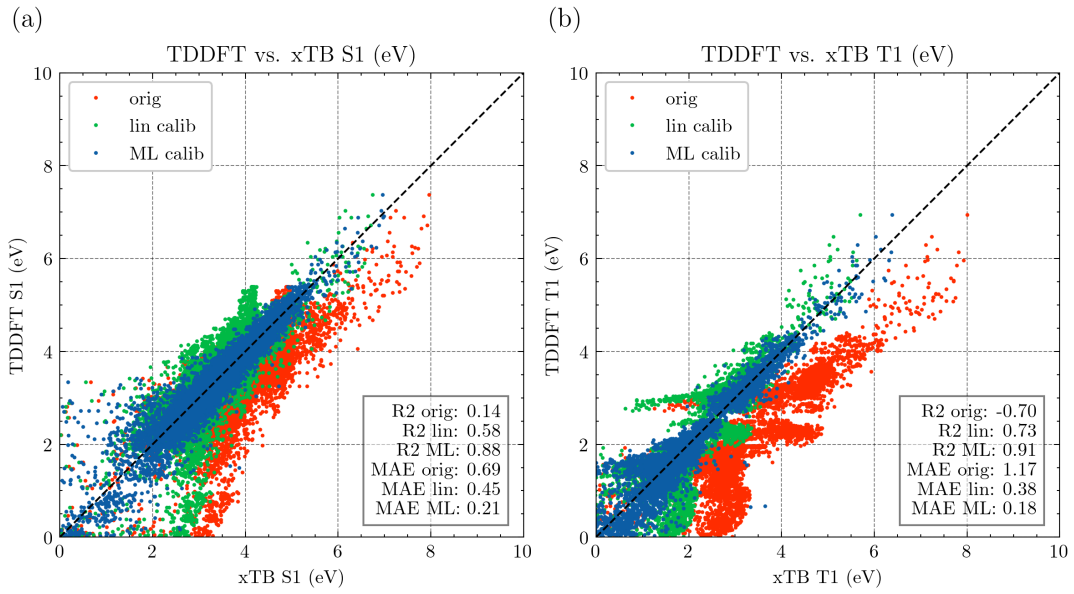


FIGURE 4.17: Plots of xTB calibration of the QM-symex-10k dataset for (a) S1 and (b) T1 energies. Red dots are original data with no calibration, green dots are linearly calibrated data, and blue dots are calibrated with ML. 10-fold cross-validation was conducted, meaning all data points shown are test points predicted by an ML model trained on the other 90% of data. Inlaid box shows quantitative measurements of accuracy for original, linearly calibrated, and ML calibrated data.

As seen in the two subsections above, when testing the ML model on subsets of the dataset in question, the ML model performs exceedingly well. However, there is the possibility that the ML model only performs well because the datasets are homogeneous, so similar molecules to those in the test set are included in the training set. To evaluate the broad applicability of our methodology, we use a blind test set of molecules not included in either of the datasets above.

4.3.3 Blind tests

MOPSSAM 143

An overarching ML model was created by combining the 10k SCOP-PCQC molecules with the 10k QM-symex-10k molecules into a 20k molecule training dataset. A 10-fold cross-validation ML model was trained with these 20k molecules and tested on the 143 molecules in MOPSSAM. As seen in Figure 4.18, the ML calibrated xTB-sTDA data matches TD-DFT values better than the linearly calibrated data. While the data are sparse, there are a few regions where the improvement is clearly visible, for example between 4.5-6.5 eV for sTDA S1, where the linear calibration over-corrects while the ML model performs better. In contrast, for low sTDA S1 energies, the linear calibration under-corrects, with the calibrated values very close to

the original values, while the ML model is more flexible. The MAE of ML-calibrated xTB-sTDA is 0.163 eV while the RMSE (not listed) is 0.24 eV for S1 energies. For T1 energies, while the ML model does outperform the linear calibration with the MAE measurement, the R2 is very similar for both techniques. This is likely because xTB nearly always over-predicts the T1 energy, so calibrating it only requires shifting in one direction, which makes linear calibration well-suited for the task. For S1 energies, there are both instances of over- and under-prediction, which makes linear calibration less applicable and motivates the need for an ML model.

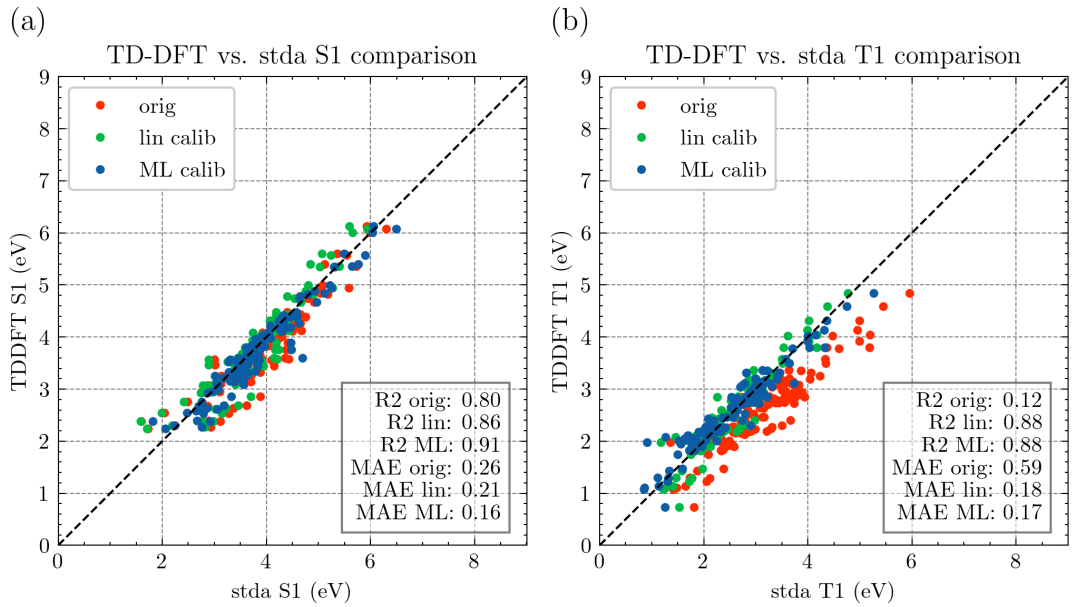


FIGURE 4.18: Plot of xTB calibration of the 143 MOPSSAM molecules for (a) S1 and (b) T1 energies. Red dots are original data with no calibration, green dots are linearly calibrated data, and blue dots are calibrated with ML. Training data was the 20k molecules in SCOP-PCQC + QM-symex-10k, and test data was the 143 molecules shown here. Inlaid boxes show quantitative measurements of accuracy for original, linearly calibrated, and ML calibrated data.

To improve these results, several expanded ML models were tested, as discussed in Section 4.2.2. First is the low-S1 PCQC expansion, choosing 2.5k molecules with low S1 energies to supplement the existing 20k training set. The second expansion uses the AL S1 cycle 1 molecules (64k) added to the 20k training set. The third expansion uses AL T1 cycle 1 molecules (107k) added to the 20k training set. The fourth expansion uses both T1 and S1 AL cycle 1 molecules (171k) plus the initial 20k training set. The fifth expansion uses all QM-symex molecules (120k) plus the initial 20k training set. The final expansion uses both AL cycle 1 molecules plus all QM-symex molecules (280k), added to the initial 20k training set. The figures showing the ML calibration plotted against original data and linear calibration are in Appendix Section A.5. Table 4.2 shows the results of all of these training sets tested on MOPSSAM 143.

As seen in this table, adding the low S1 molecules reduces the predictive power of the ML model for the S1 energies but improves the T1 model very slightly. Thus,

TABLE 4.2: Accuracy of various ML model training set expansions on predicting MOPSSAM excited state energies

Model	Training Set Size	S1 R2	T1 R2	S1 MAE	T1 MAE
No calibration	N/A	0.80	0.12	0.26	0.59
Linear calibration	143	0.86	0.88	0.21	0.18
SCOP-PCQC + QM-sym-10k	18,965	0.91	0.88	0.16	0.17
SCOP-PCQC + QM-sym-10k + LowS1	21,440	0.80	0.87	0.21	0.16
SCOP-PCQC + QM-sym-10k + ALS1	77,910	0.90	0.90	0.18	0.15
SCOP-PCQC + QM-sym-10k + ALT1	126,065	0.86	0.90	0.19	0.13
SCOP-PCQC + QM-sym-10k + ALS1 + ALT1	181,629	0.88	0.91	0.18	0.13
SCOP-PCQC + QM-sym-10k + QM-symex	138,435	0.91	0.90	0.17	0.14
SCOP-PCQC + QM-sym-10k + ALS1 + ALT1 + QM-symex	301,099	0.90	0.93	0.17	0.13

although the ML model was trained on a larger dataset, we can infer the composition was inadequate for predicting excited state energies. Similarly, adding the AL S1 data failed to improve the ML model's S1 predictions, while the T1 predictions did improve slightly. Adding the AL T1 molecules vastly improved the predictive performance for T1 energies, reducing the MAE from 0.17 to 0.13 eV, while the S1 prediction became slightly worse from 0.16 to 0.19 eV. Adding both AL S1 and AL T1 molecules was a decent compromise, keeping the T1 MAE low at 0.13 and the S1 MAE at 0.18. Transitioning to the QM-symex dataset, adding all 120k molecules improves the T1 MAE and keeps the S1 MAE essentially the same. Finally, adding all expansions (AL S1, AL T1, and QM-symex) showed the lowest T1 MAE of 0.13 eV, but had limited impact to S1 MAE (0.17 eV).

From this data, it is clear that adding additional molecules helps improve the T1 predictions, but this is not reflected in S1 predictions. The reasoning is not immediately clear, but there are a few possible explanations. First is that the T1 energies are more or less linearly calibrated, but molecules lie on different lines, so adding additional molecules may help refine this calibration. In contrast, S1 energies may

be distributed around the $x = y$ line uniformly and adding additional points may not help. Another explanation is that the initial 20k training set may already have enough similar molecules to MOPSSAM 143 to have high predictive power, and adding additional molecules may unnecessarily obfuscate the model. The additional molecules may help cover more of the global chemical space but not necessarily improve performance for the 143 molecules considered here. Therefore, further analysis is required to determine which model is the best. 2 models, SCOP-PCQC + QM-sym-10k and SCOP-PCQC + QM-sym-10k + ALS1 + ALT1 + QM-symex, are chosen for further analysis, and are referred to as xTB-ML-20k and xTB-ML-300k, based on their training size, for simplicity. The first alternative test set used is MOPSSAM 1000.

MOPSSAM 1000

Because these 143 molecules are sparse, to test our ML model further we sample 1k molecules from the 250k molecules in MOPSSAM's test set. S1 and T1 values were calculated with TD-DFT and xTB-sTDA. The xTB values were then calibrated with a linear model trained with the 143 molecules in MOPSSAM 143, to be consistent with Wilbraham et al.'s methodology.[112] ML-calibrated xTB values were calculated using both the xTB-ML-20k and xTB-ML-300k models. Figures 4.19 and 4.20 show the results of this test.

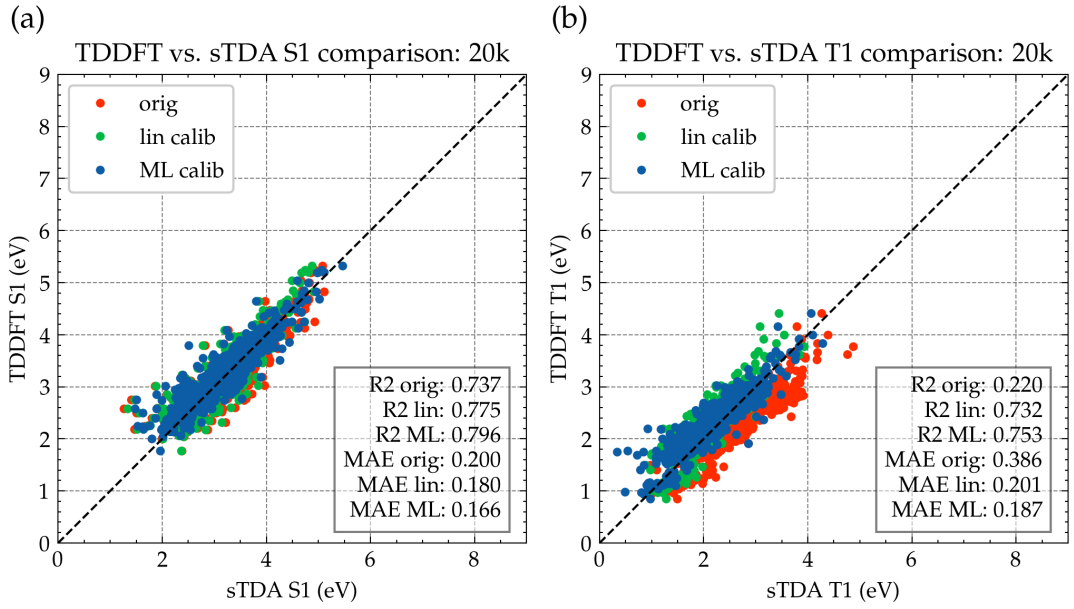


FIGURE 4.19: Plot of xTB calibration of 1k randomly selected MOPSSAM molecules for (a) S1 and (b) T1 energies. Trained on 20k molecules in SCOP-PCQC + QM-symex-10k and tested on MOPSSAM 1000.

As seen, both ML models outperform linear calibration. xTB-ML-300k does slightly better for S1 prediction, and much better for T1 prediction. We again visually see the linear shift required for T1 calibration, while S1 calibration is not as clear. The 300k model specifically does better at lower energies (<2 eV), effectively

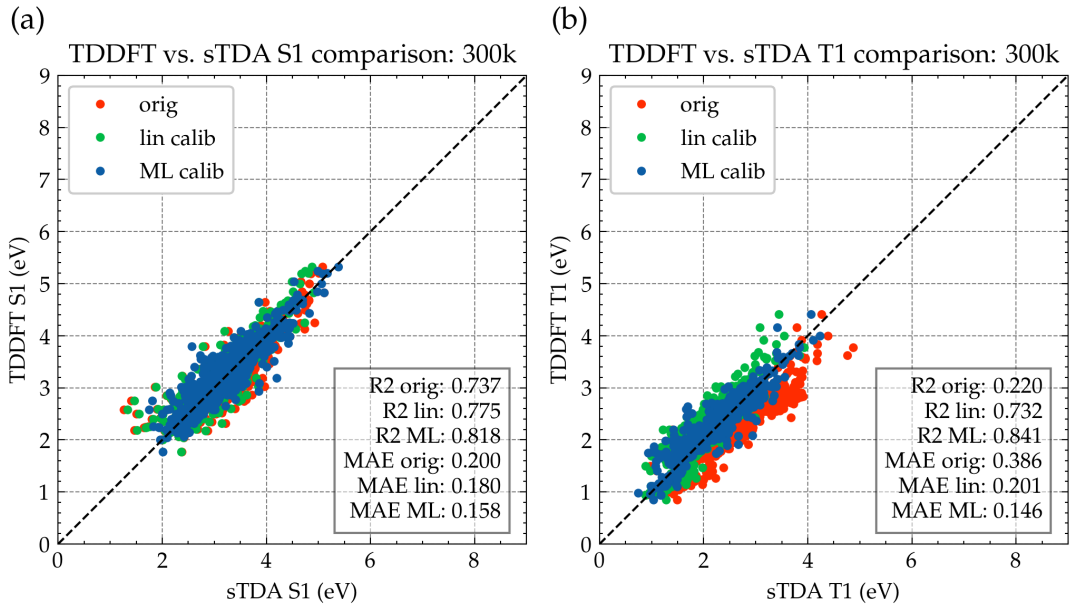


FIGURE 4.20: Plot of xTB calibration of 1k randomly selected MOPSSAM molecules for (a) S1 and (b) T1 energies. Trained on 300k molecules in SCOP-PCQC + QM-symex-10k + AL S1 + AL T1 + QM-symex and tested on MOPSSAM 1000.

removing outliers. Both ML models are able to accurately calibrate energies without over- or under-correcting, while the linear calibration again tended to over-correct high energies and under-correct low energies.

It seems that xTB-ML-300k outperforms the 20k version, but to confirm this it is useful to test on a different dataset.

INDT

To test the broader applicability of the xTB-ML models, we use the INDT blind test dataset. 1k molecules were sampled from this dataset for ease of viewing. Figures 4.21 and 4.22 show the results of this test.

For this dataset, xTB-ML-20k vastly outperforms xTB-ML-300k, and both definitively outperform linear calibration. The excited state energies of these molecules are similar, so the linear model is unable to discern the nuances in calibration. xTB-ML-20k is able to shift the data so that it is more tightly bound around the $x = y$ line. In contrast, the xTB-ML-300k model is unable to shift sufficiently – this difference is most clearly seen on the right side of S1 calibration, where the blue points are overlapping the green and red points, indicating insufficient calibration. The difference in performance could be because there are more INDT-like molecules in the 20k training set, which allows xTB-ML-20k to calibrate more accurately. This is promising as it indicates the training set generation technique for 20k is comprehensive and includes molecules potentially of interest to TTA and SF. Overall, both models outperform linear calibration and are able to keep the MAE at or below 0.2 eV, with the xTB-ML-20k model performing its best at 0.13 eV.

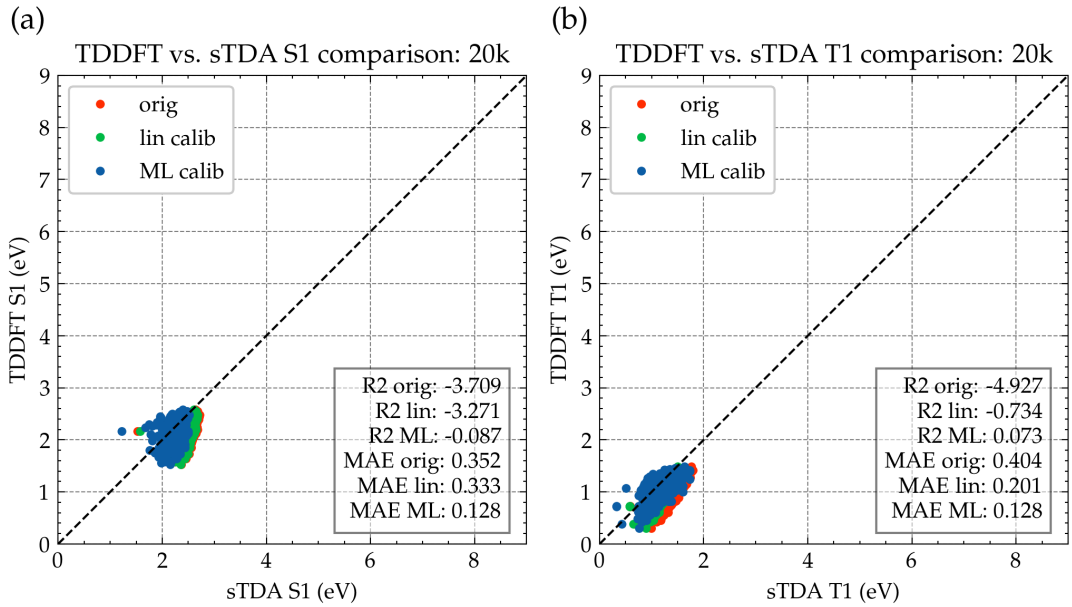


FIGURE 4.21: Plot of xTB calibration of 1k randomly selected INDT molecules for (a) S1 and (b) T1 energies. Trained on 20k molecules in SCOP-PCQC + QM-symex-10k and tested on INDT.

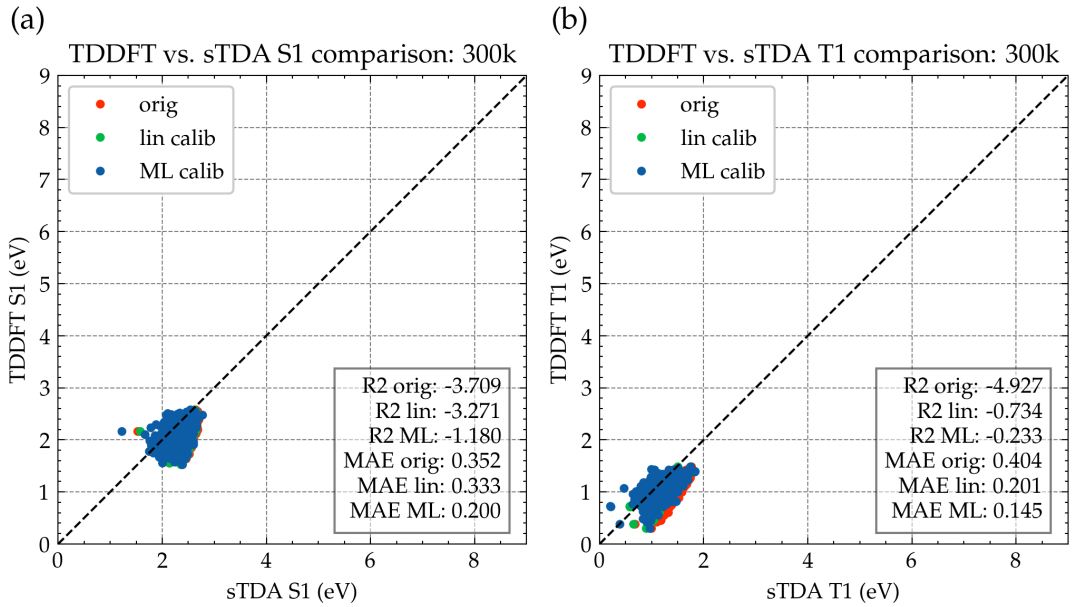


FIGURE 4.22: Plot of xTB calibration of 1k randomly selected INDT molecules for (a) S1 and (b) T1 energies. Trained on 300k molecules in SCOP-PCQC + QM-symex-10k + AL S1 + AL T1 + QM-symex and tested on INDT.

As seen, there are datasets where xTB-ML-20k performs better, and datasets where xTB-ML-300k performs better. On average, compiling MAEs for all 6 datasets considered (MOPSSAM 143, MOPSSAM 1000, and INDT, for both S1 and T1), the 20k model has an MAE of 0.157 eV while the 300k model has an MAE of 0.158 eV. Given this small difference, it is difficult to choose a model for further study. Considering the 20k model performed better for the external INDT dataset, and as it

requires less training data and therefore would be more applicable to future studies employing ML calibration of datasets, this model is used in the following sections, and is referred to as xTB-ML.

Overall, in this section we have shown the greater accuracy of our ML-calibrated xTB-sTDA results in comparison to raw or linearly calibrated data, when compared to a TD-DFT benchmark. This improvement is first evident in test sets composed of subsets of the training datasets, but is shown to be upheld with external test sets as well. Notably, the blind test sets used had different TD-DFT settings (different initial 3D structure generation, different basis set, inclusion of solvation model) but the ML calibration still improved the accuracy of results.

4.3.4 Coupled cluster calibration

The natural next question is whether the xTB-ML model can be applied to different TD-DFT functionals or different computational techniques altogether such as CC2 or ZINDO. To test this, we ran xTB-ML on QM8,[83,91] which includes S1 excited state energies at the PBE0/def2-SVP, PBE0/def2-TZVP, and CAM-B3LYP/def2-TZVP levels of theory for TD-DFT as well as the RI-CC2/def2-TZVP coupled-cluster level of theory, and QM7b,[84,90] which includes S1 energies at the ZINDO level of theory. Unfortunately, as seen in Appendix Section A.6, xTB-ML does not improve accuracy. Therefore, it can be assumed that xTB-ML would best be applied when calibrating xTB results against TD-DFT calculations with the B3LYP functional.

However, it is possible to apply the same calibration methodology presented in this work to calibrate xTB against CC2, similar to how previously xTB was calibrated against TD-DFT. For this, we use the CC2 values compiled in QM8, randomly sampling 10k molecules as the training set and using the other 11.5k as the test set. This time, because of the smaller dataset, we use 20-fold cross-validation with 95%/5% train/validation splits. This helps ensure most of the data is used in training. The new model is termed xTB-CC-ML to distinguish it from the previously generated xTB-ML model.

Figure 4.23(a) shows the results of the comparison, with the inlaid box displaying measurements of accuracy for both methods. As seen, adding the ML calibration to xTB results vastly improves results, reducing the MAE by 66%. For comparison, Figure 4.23(b) shows the results of TD-DFT calculations on the same test set of molecules. As seen, xTB-CC-ML has higher accuracy than TD-DFT calculations for the 11.5k test set, using either R2 or MAE as the metric.

To test the impact of training size on accuracy, 8 different ML models were generated with training sizes ranging from 100 to 15,000. The models were then predicted on the remaining molecules in QM8 not used in the training set. The MAE of the test set (against CC2 values) was compared to the MAE of PBE0/def2-TZVP and CAM-B3LYP/def2-TZVP, as shown in Figure 4.24.

As seen, a training size of approximately 750 molecules allows xTB to achieve similar accuracy to PBE0. It is more difficult to match CAM-B3LYP's accuracy, but

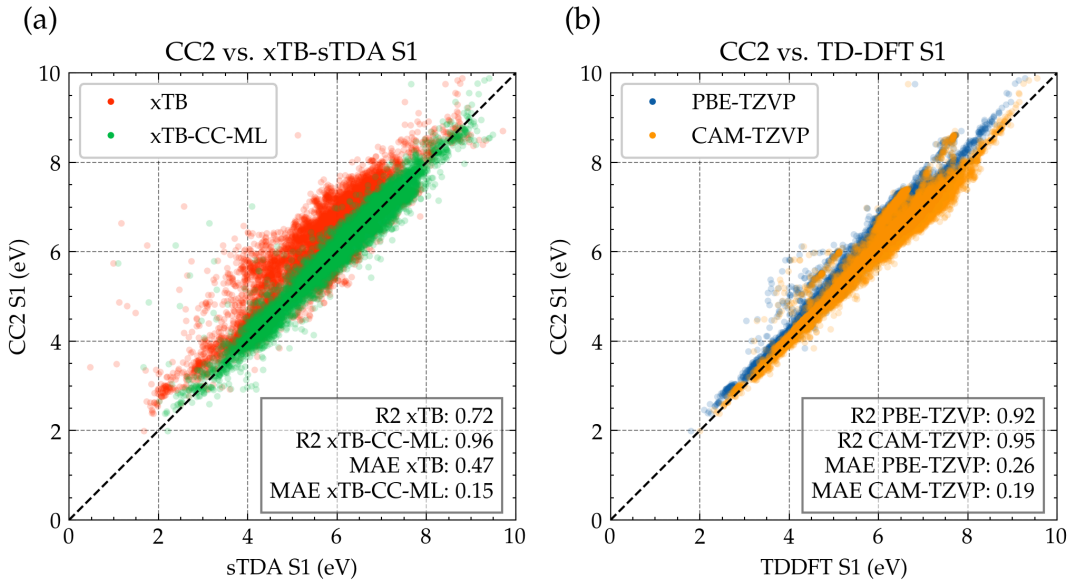


FIGURE 4.23: (a) Plot of xTB calibration against CC2 with 10k training set and 11.5k test set (shown). Red dots indicate original xTB calculations while green dots indicate calibrated xTB data. (b) Plot of TD-DFT calculated values against CC2 values, for accuracy comparison. Black dashed line in both plots indicates $x = y$ line.

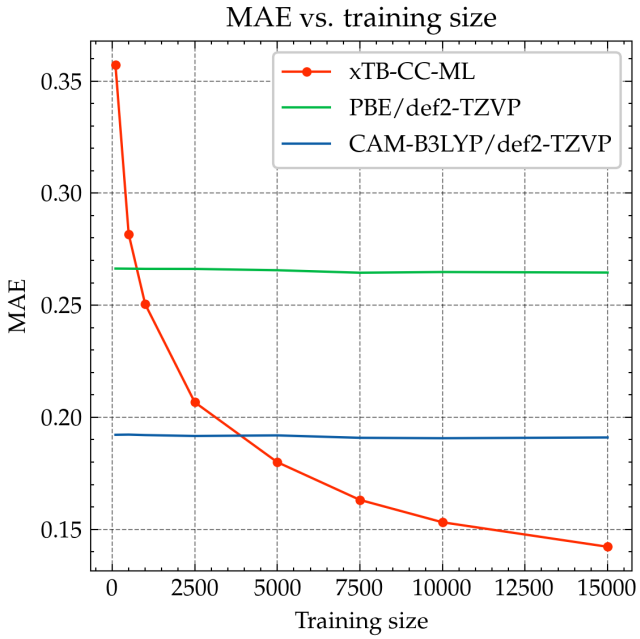


FIGURE 4.24: (a) Plot of xTB calibration against CC2 with 10k training set and 11.5k test set (shown). Red dots indicate original xTB calculations while green dots indicate calibrated xTB data. (b) Plot of TD-DFT calculated values against CC2 values, for accuracy comparison. Black dashed line in both plots indicates $x = y$ line.

this is achieved at a training size of around 3750. At the largest training size considered (15k), the xTB-CC-ML calibrated molecules vastly outperform both TD-DFT results, with a 47% reduction in MAE for PBE0 and 27% reduction in MAE for CAM-B3LYP.

These are incredibly promising results, showing machine learning can help high-throughput techniques such as xTB-sTDA compete with high-accuracy methods such as TD-DFT. However, the xTB-CC-ML model may not be as generalizable as xTB-ML, due to the smaller (10k), less diverse (only small molecules with up to 8 heavy atoms) training set. For this reason, xTB-ML will be used in the following sections instead of xTB-CC-ML. Regardless, xTB-CC-ML serves as a interesting proof of concept that can be expanded further in the future, perhaps with additional CC2 calculations on more diverse molecules.

Regardless, now that we have a functioning ML model for improving the results output by xTB (xTB-ML), we can use the model for various applications. The following sections will apply the model to calculate excited states in large databases and map inaccuracies of xTB-sTDA in chemical space, after a brief detour to discuss the benefits of xTB-ML versus direct ML.

4.3.5 Comparison to direct ML

Before continuing with applying the xTB-ML model, we first must prove such a calibration model is more accurate than a direct ML model when predicting excited state energies. The reasoning is that xTB gets the excited state energy in the right region, so the load on the ML model is lower as it simply needs to nudge the value in a certain direction. To test this theory, we generate an ML model trained on the same 20k molecules as above but instead directly set the energy as the objective, and test the model on MOPSSAM 143. Figure 4.25 shows the results.

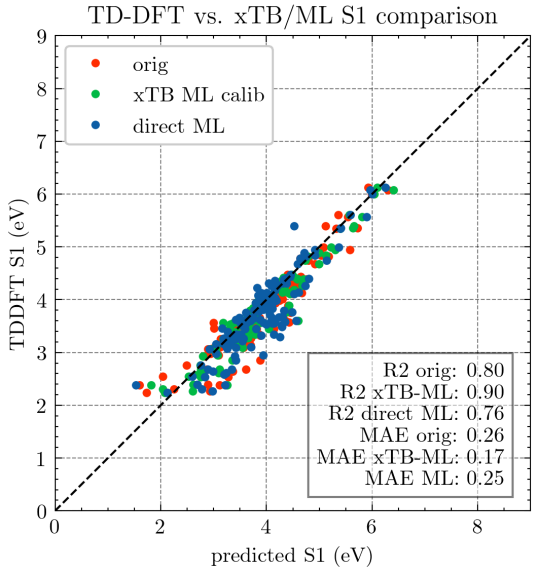


FIGURE 4.25: Comparison of ML-calibrated xTB results vs. directly predicted energies with ML. Red dots are original data with no calibration, green dots are ML calibrated xTB-sTDA data, and blue dots are directly generated with ML. Training data was the 20k molecules in SCOP-PCQC + QM-symex-10k, and test data was the 143 molecules shown here. Inlaid box shows quantitative measurements of accuracy for original xTB-sTDA, ML calibrated xTB-sTDA, and direct ML data.

As seen, the performance of the directly trained ML model is vastly lower than the ML-calibrated xTB data. The ML model performs worse than xTB itself in some metrics, and the ML model performs worse than the linear calibration model in all metrics. Thus, calibrating xTB with ML gives much higher accuracy than using ML to directly predict energies.

The one benefit of direct ML prediction is the time savings. Both xTB-ML and direct ML require training an ML model – for the 20k training set presented in this work, training took approximately two hours. Direct ML has a much faster prediction, however – it predicted the S1 energies of the 143 molecules above in less than 1 minute, and the S1 properties of all 250k molecules considered in MOPSSAM in 1 hour and 12 minutes (on a workstation with 24 CPUs, 188G memory). For comparison, xTB-ML computes properties of approximately 1500 molecules per hour when parallelized over 4 nodes with 24 CPUs and 250G memory each. Despite direct ML being several orders of magnitude faster than xTB-ML, the low accuracy of the method reduces its efficacy.

Now that we have shown the xTB-ML model outperforms direct ML models, we can move on to using our xTB-ML model. The first section will be on applying our model to the 250k molecules in MOPSSAM.

4.3.6 Calculating excited state energies for 250k molecules

We can now apply our xTB-ML model to the 250k molecules considered in Wilbraham et al.[112] to see how different the S1 and T1 energies would be. The results are shown in Figure 4.26. For S1, the ML calibrated data seems to be generally centered around the original data, and the linear calibration is minimal until higher energies. For most molecules, there is approximately a ± 0.5 eV difference between linearly calibrated and ML calibrated values, with the deviation decreasing at lower energies. However, for higher energies, the linear calibration starts to underestimate in comparison to the ML calibration. This is reflected in Figure 4.18, where the linear calibration shifts the data too far to the left for high energies. The T1 calibration is more significantly different between ML and linear calibrations. At low energies, the linear model does not change the energies much, while the ML model lowers the energy. In contrast, at high energies, the linear model decreases the energy by as much as 1 eV, while the ML model is more conservative in its calibration. Both ML and linear calibration down-shift the calculated energy for most molecules, which is in line with expectations. Note that because TD-DFT data was not calculated for these 250k molecules, we cannot compare the calibration to ground truth, but based on the metrics presented in Section 4.3, it is likely the ML-calibrated values are more accurate.

Now that we have both S1 and T1 energies calculated for 250k molecules with xTB-ML, we can identify potential sensitizers and emitters. We can use the following figures of merit (FOMs) to evaluate whether a molecule is a suitable candidate based

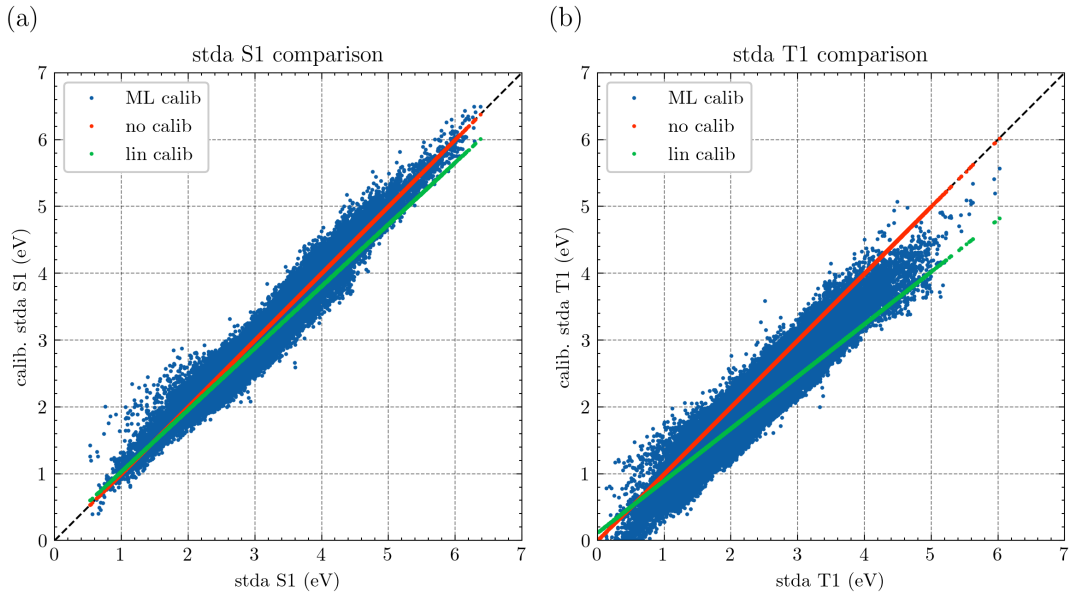


FIGURE 4.26: Plot of 250k molecules showing difference in calibrated (a) S1 and (b) T1 with ML model compared to linear model. Red dots are without calibration, green dots are with linear calibration, and blue dots are with ML calibration.

on its excited state energies.

$$\text{FOM}_{sens} = \begin{cases} 0 & T1 > S1 \\ e^{-\left|1-\frac{T1}{S1}\right|} + e^{-|1.1-S1|} & S1 \geq T1 \end{cases}$$

$$\text{FOM}_{emit} = \begin{cases} 0 & S1 > 2T1 \\ e^{-\left|2-\frac{S1}{T1}\right|} + e^{-|2-S1|} & S1 \leq 2T1 \end{cases}$$

The first check is if the energies are invalid, i.e. if the T1 of the sensitizer is greater than the S1, or if the S1 of the emitter is more than twice the T1. If these are true, then the molecule is discarded and no calculation is necessary. If the molecule is indeed valid, then there are two terms that compose the FOM. The first term ensures the ratios are as close as possible to ideal – i.e. there is minimal loss in energy. The second term in each FOM ensures the singlet energy of the sensitizer and emitter match the target properties. This term is flexible depending on the application – for this work 1.1 eV was chosen as the excitation energy of the sensitizer, which would function as a near-IR absorber, and 2 eV was chosen as the emission energy, creating a visible emitter. Note that the exponential terms constrain each component of the FOM between 0 and 1, so the maximum FOM is 2. Figure 4.27 shows the results of screening molecules for potential sensitizers and emitters.

We have therefore used xTB-ML to make quick, relatively accurate calculations for S1 and T1 energies, and have used the results to screen for potential sensitizers and emitters. While this is useful for screening existing databases for suitable molecules, generating new molecules for TTA would also be useful, as detailed in the following section.

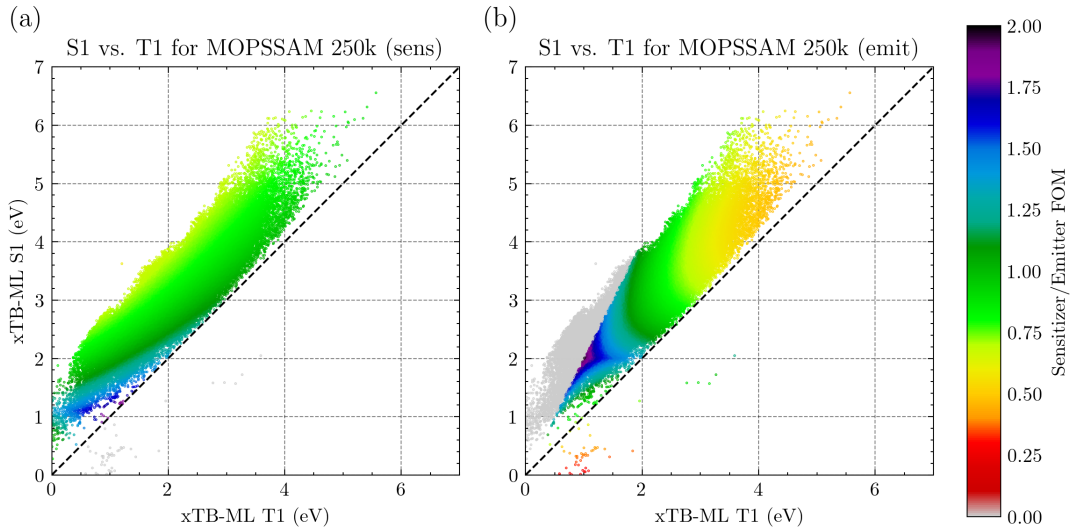


FIGURE 4.27: Plot of 250k molecules showing S1 and T1 energies, colored with FOM for (a) sensitizers and (b) emitters. Target properties include correct ratio of T1/S1 and S1 near certain values. 0 molecules have sensitizer FOM > 1.9, but 60 molecules have emitter FOM > 1.9. SMILES, S1, T1, and FOM for molecules are available on GitHub.[142]

4.3.7 Mapping inaccuracies in chemical space

Since our ML model predicts the error in xTB-sTDA, an interesting application is to map the error in S1 and T1 calculations in a global chemical space, to see if there are some areas where xTB systematically over- or under-estimates, or areas where xTB is projected to be fairly accurate. The first requirement is to generate a global chemical map. We used UMAP[141] to embed the high-dimensional molecular data into 2 dimensions, using the Jaccard-Tanimoto similarity between Morgan fingerprints of molecules for proximity. We then colored the global chemical space map in 3 different ways, as shown in Figure 4.28.

First, we used HDBSCAN[155] to cluster the molecules based on proximity, as shown in Figure 4.28(c). HDBSCAN is a soft clustering, not creating distinct categories but instead giving molecules a rating between 0 and 10 (or -1 for no cluster, as approximately 1/3 of the molecules were unable to be clustered). We can see that HDBSCAN effectively clusters molecules in space, with most molecules in close proximity included in the same cluster. Some of the clusters themselves are spread out across space, such as the purple cluster that includes many molecules along the edge of the global space. Note that this is a dataset-agnostic clustering, as the clustering algorithm only sees molecular information and no labelled data. More details about the HDBSCAN algorithm can be found in the paper by Campello et al.[155] and website.[156]

For the next two plots (Figure 4.28(a) and (b)), we used our ML model to predict the error in xTB-sTDA. Here, the error is defined as:

$$\text{error} = \text{true energy} - \text{xTB-sTDA energy} \tag{4.1}$$

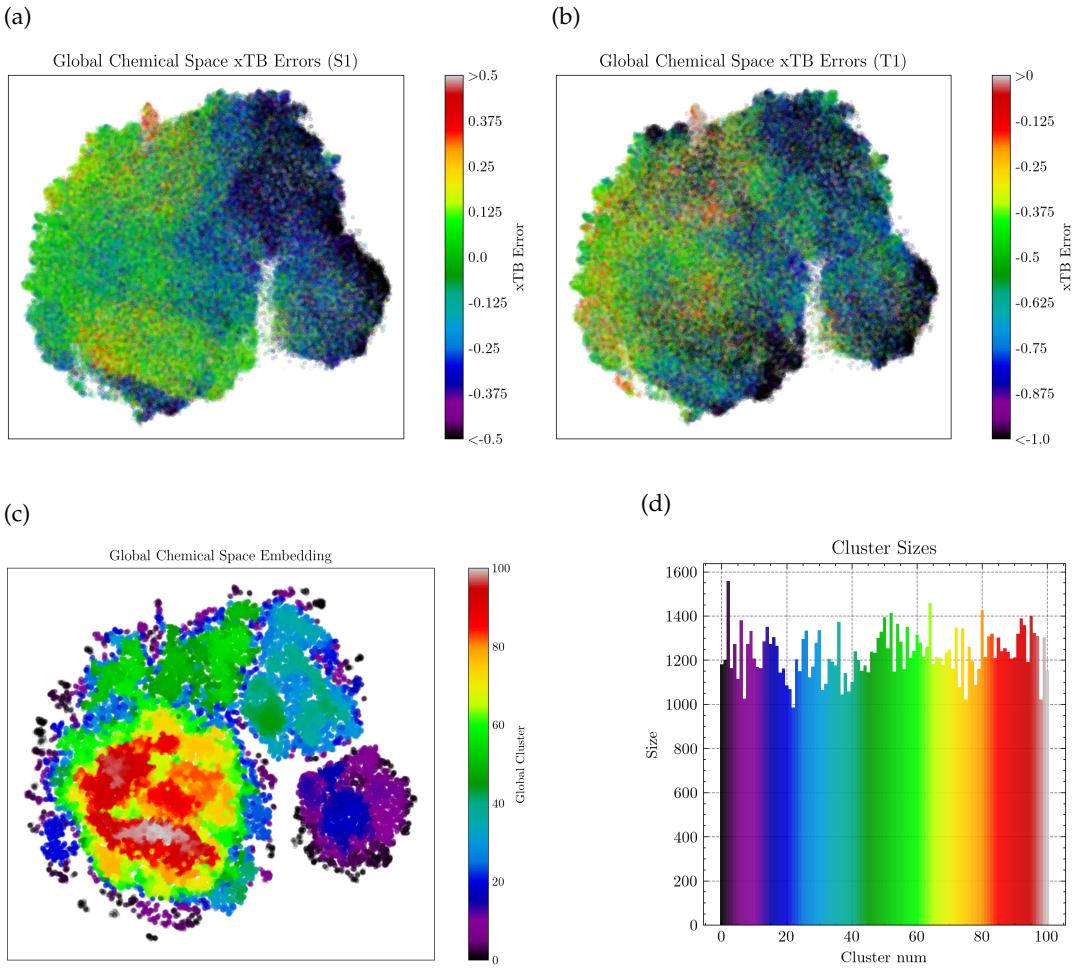


FIGURE 4.28: Global chemical space maps. Plots of xTB (a) S1 and (b) T1 errors in global chemical space. (c) Clustering of molecules in global chemical space. (d) Number of molecules per cluster in global chemical space.

so a negative error implies xTB-sTDA is over-predicting the excited state energy. As seen in Figure 4.28(a), there are distinct regions where xTB-sTDA over-predicts S1 (right side), regions where xTB-sTDA has reasonable accuracy (top left and center), and regions where it under-predicts (left and top). In general, most molecules are within $\pm$0.5eV error.

In contrast, for the T1 energy, xTB-sTDA over-predicts for almost all molecules, as seen in Figure 4.28(b). Note that the scale in this plot is shifted from -0.5–0.5 eV (as in S1) to 0– -1.0 eV, to make the distribution of errors clearer. Only a few scattered molecules are under-predicted by xTB-sTDA and are colored red, and all other molecules are over-predicted. Similar to S1, xTB-sTDA over-predicts T1 for most molecules on the right side, and gets reasonable accuracy on molecules in the middle and top left. In contrast to S1, T1 is also over-predicted on a cluster of molecules on the bottom left.

The next natural question is whether each cluster as defined by HDBSCAN has a particular error associated with it. For example, it seems that xTB-sTDA does a relatively good job for the red cluster, but over-predicts energies for molecules in blue and green clusters. Although HDBSCAN is a soft clustering, we can categorize molecules into 100 distinct clusters based on the number assigned to them, as well as 2 additional clusters (1 each for unclustered molecules and for outliers). Figure 4.29 quantifies the mean errors for S1 and T1 energies for each cluster.

Subplots (a) and (b) show the mean errors (ME) of S1 and T1, and we can see our qualitative observation of orange/red clusters having low error and blue/green clusters having high error is empirically upheld. We also note that a few dark blue clusters seem to have low errors, for both S1 and T1. However, visually inspecting the molecules in the dark blue clusters, it seems some have high negative and positive error. To more accurately depict each clusters error, it is useful to take the absolute value of the error before averaging them (MAE). This is shown in plots (c) and (d) in Figure 4.29. With these plots, we again see that the blue/green clustered molecules have high error, while red/orange/yellow clusters have low error. There is still one dark blue cluster with low error, but this could be due to chance, as the cluster picked molecules with low error, since most other dark blue clusters have high error.

For broader applications, knowing the error expected for each cluster is useful if the location and cluster categorization of a specific molecule in global chemical space is known. Oftentimes, this is not known, or would require significant computation. It would be ideal to know the properties of molecules with low predicted error, to have greater confidence in xTB-sTDA calculations, or those with high error, to know to use the ML model or consider other computational techniques. The following subsection discusses this further.

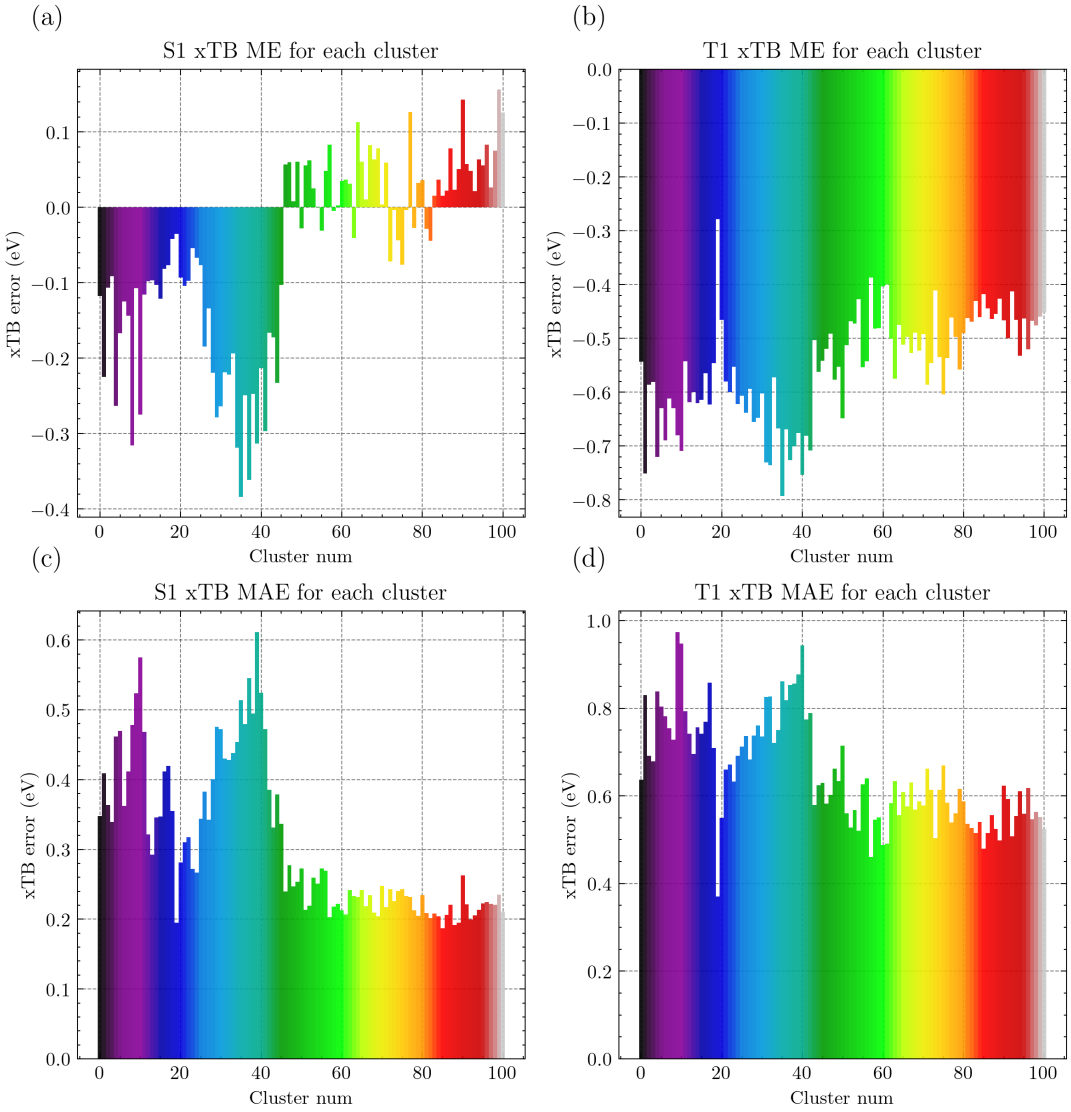


FIGURE 4.29: Mean errors for S1 and T1 energies of molecules in global chemical space. (a) Mean S1 error, (b) mean T1 error, (c) mean absolute S11 error, and (d) mean absolute T1 error. Note that clusters 0-9 are relevant, cluster -1 includes unclustered molecules while cluster 10 includes outliers. "Absolute" error takes the absolute value of errors before averaging them.

Characteristics of molecules with low/high error

When running xTB-sTDA, it would be useful to have some idea as to whether it will return accurate results. While clustering as discussed previously is an option, it would be more beneficial to have some chemical intuition of accuracy based on the molecular structure. To this end, this subsection identifies substructures that are more likely to be present in low-error or high-error molecules.

We first use our ML model generated above to predict the S1 and T1 error between xTB and TDDFT for randomly subsampled 1M molecules from PCQC. We then categorize the molecules based on the predicted error as follows:

$$\begin{aligned} \text{Cat}_{S1} &= \begin{cases} \text{Low} & |S1_{err}| < 0.05 \\ \text{High}_{\text{Under}} & S1_{err} > 0.5 \\ \text{High}_{\text{Over}} & S1_{err} < -0.5 \end{cases} \\ \text{Cat}_{T1} &= \begin{cases} \text{Low} & |T1_{err}| < 0.05 \\ \text{High}_{\text{Under}} & T1_{err} > 0 \\ \text{High}_{\text{Over}} & T1_{err} < -1.0 \end{cases} \end{aligned} \tag{4.2}$$

where "under" refers to xTB underestimating the energy while "over" refers to overestimating (note the error definition in Equation 4.1). For both T1 and S1 error, we define low error as $<\pm 0.05$ eV. However, for defining high error, for T1 we shift the bounds down by 0.5 to reflect the distribution of errors, as seen in Figure 4.28(c). The percentage of molecules in each category is shown in Table 4.3.

TABLE 4.3: Percentage of molecules in each error category, with categories defined in Equation 4.2

	Low	**High, Under**	**High, Over**	**Average**
S1	15	1.5	11	72.5
T1	2.4	4.3	17	76.3

where "average" denotes molecules with uncategorized error. As seen, for both energies, the majority of molecules are uncategorized. However, we can conduct substructure analysis on the low and high overestimation categories (ignoring high underestimation due to low fraction of molecules) to know when to trust the xTB-sTDA results, or when to expect exceptionally high errors.

We use molZ[157] to analyze which substructures are over-represented in each category. The results of this substructure analysis are shown in Figures 4.30 for low-error molecules and 4.31 for high-error molecules.

From these plots, a few patterns become evident. Low error molecules are more likely to be aromatic, potentially with sequential attached rings, for both S1 and T1.

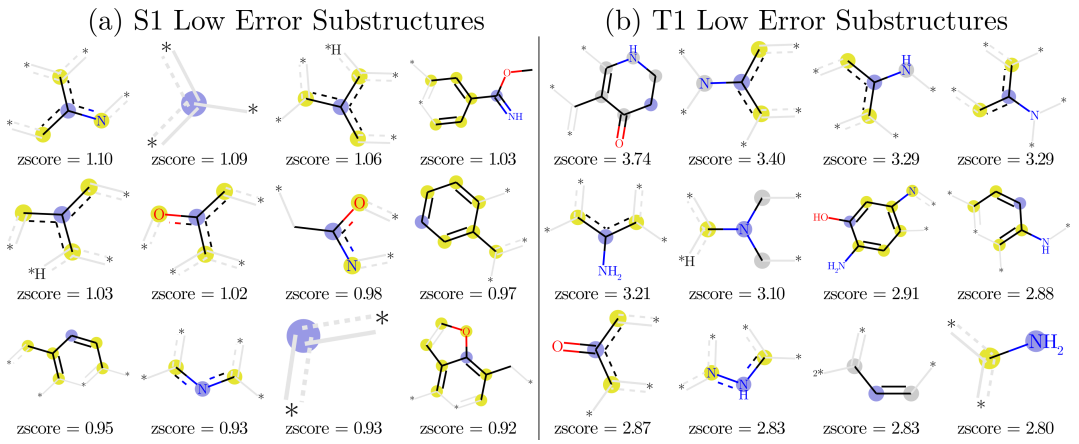


FIGURE 4.30: Grid of molecular substructures over-represented in molecules with low error as predicted by the ML model, for (a) S1 error and (b) T1 error. According to RDKit, blue atoms are the center atoms, yellow atoms are aromatic atoms, dark gray atoms are aliphatic ring atoms, and light gray atoms/bonds are connectivity invariants.

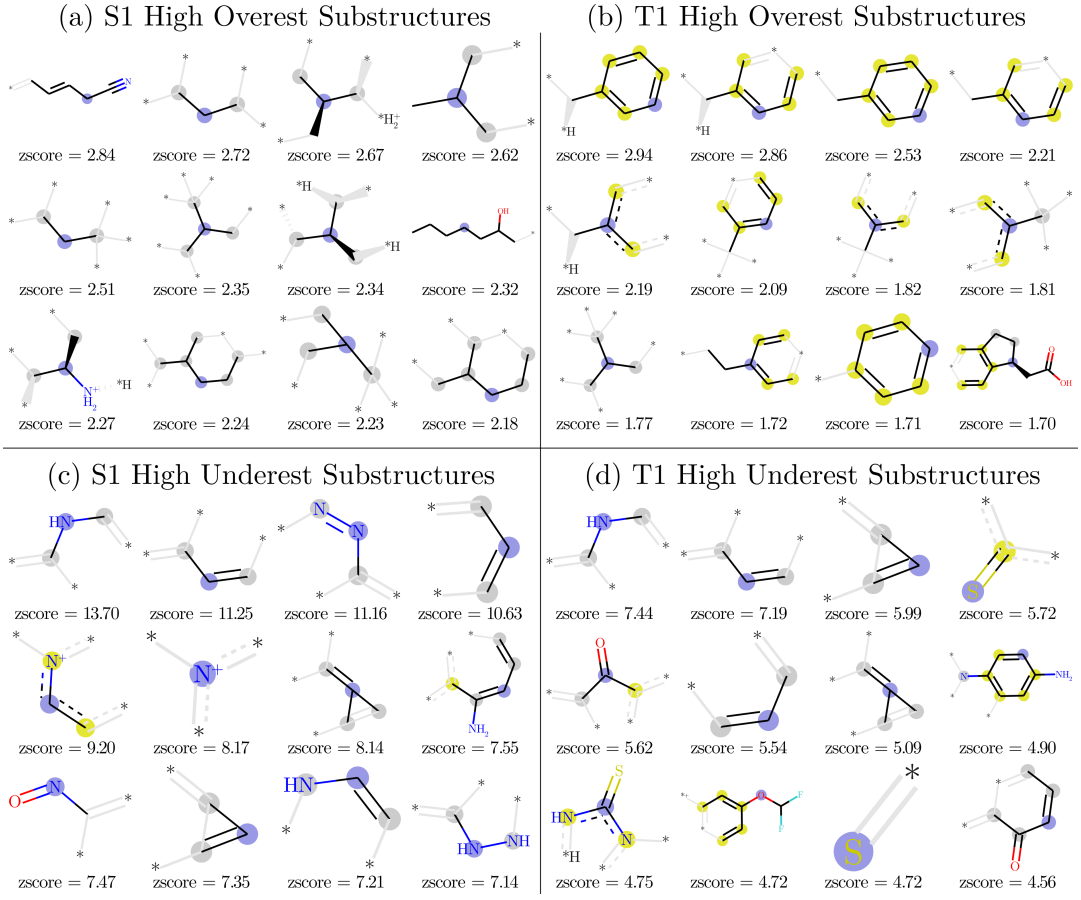


FIGURE 4.31: Grid of molecular substructures over-represented in molecules with high error as predicted by the ML model, for (a) S1 overestimation error, (b) T1 overestimation error, (c) S1 underestimation error, and (d) T1 underestimation error. According to RDKit, blue atoms are the center atoms, yellow atoms are aromatic atoms, dark gray atoms are aliphatic ring atoms, and light gray atoms/bonds are connectivity invariants.

In contrast, S1 high overestimation, S1 high underestimation, and T1 high underestimation molecules are likely to be not aromatic, with some unconventional molecular structures included in these groups. There are some aromatic substructures in the T1 overestimation molecules, but they are attached to the bulk structure with a rotatable bond. This overestimation could be a result of the 3D structure generation, since only limited conformer analysis is conducted and potentially the lowest energy conformer was not achieved. To clarify the effect of this versus an inherent inaccuracy in the excited state energy calculation of xTB-sTDA, a more intensive conformer search could be an avenue of future work.

Overall, these predictions can be used as guides to develop chemical intuition for the accuracy of the xTB-sTDA methodology in calculating excited state energies.

4.4 Conclusions and Future Work

In this chapter, we have presented a methodology for calibrating a high-throughput computational chemistry technique (xTB-sTDA) against a high-accuracy one (TD-DFT) using ML. We first decided on Chemprop's MPNN as the model of choice, then generated a training set using literature scraping of relevant molecules from abstracts (SCOP-PCQC) and an existing excited state database (QM-symex-10k). We also generated a blind test set from a previously published study conducting linear calibration of xTB-sTDA results (MOPSSAM). We then compared ML models generated from the small training set to various other expanded datasets, but found the small 20k-molecule training set performed similarly to if not better than the largest training sets considered (300k). In all cases, the ML calibration model (xTB-ML) outperformed linear calibration, oftentimes significantly. All xTB-ML calibrated models had MAEs of less than 0.2 eV when compared to TD-DFT, and in some cases the error was much less (~0.13 eV).

The xTB-ML model was the main result of this work. For the 3 blind test sets used in this study, the average MAE was 0.341 eV for only xTB-sTDA, 0.225 eV for linearly calibrated xTB-sTDA, and 0.150 eV for xTB-ML. If xTB-ML is used as the first step in a high-throughput screening process instead of raw xTB outputs, its low error can help ensure that all relevant molecules are selected and not weeded out, while simultaneously ensuring all selected molecules are relevant. Having a high-throughput technique with high accuracy can therefore be extremely useful for materials discovery.

After evaluating the performance of xTB-ML, we then used the model for three applications. First was comparing the xTB-ML model against directly predicted energies with ML, showing the xTB-ML model had better accuracy (0.17 eV vs. 0.25 eV MAE). Second was rapidly screening 250k molecules for suitability as NIR-TTA sensitizers and emitters. The databases had no suitable sensitizers, but 60 suitable emitters were found. Lastly was mapping inaccuracies of xTB in chemical space, using the ML model to predict errors in xTB. This was used to see which regions of

chemical space xTB has high errors in. S1 errors were small, with most molecules being within 0.5 eV. There were clear regions where xTB overpredicted S1, but only a few for underprediction. T1 energies were generally overpredicted, with most molecules being between 0 and 1 eV below TD-DFT values. Global chemical space mapping provides another method of predicting xTB error, by calculating which cluster a molecule belongs to and referencing the MAE of that cluster. Properties of low-error molecules were also evaluated, finding non-aromatic molecules are likely to have higher error.

While xTB-ML was able to improve performance of xTB-sTDA against TD-DFT values calculated with B3LYP, the model was unfortunately not generalizable to other functionals (PBE0, CAM-B3LYP) or other computational chemistry techniques (ZINDO, CC2). However, applying the same methodology used in this study to calibrate xTB-sTDA against coupled cluster values (CC2), and generating a new xTB-CC-ML model, showed calibrated xTB-sTDA values had high accuracy (0.15 eV MAE), out-performing TDDFT values calculated with PBE0 (0.26 eV MAE) and CAM-B3LYP (0.19 eV MAE). Therefore, the methodology presented here is generalizable to other calibrations.

There are a few avenues of future work to mention here. First is improving the ML model architecture. While Chemprop's MPNN outperformed other ML models, primarily due to its advanced featurization, it still only took the 2D molecular structure as input. Since the 3D structure is already output by xTB, including this information as input to the ML model would likely improve performance. Another improvement to the ML workflow would be to conduct a more intensive conformer search. While OpenBabel's `gen3D` function includes a search for 200 conformers, these may not include the lowest energy conformer, thus reducing the accuracy of the xTB portion of the workflow. It would be unfair to characterize this error as error due to xTB, since this is due to initial structure generation. Using a conformer searching tool such as CREST[158] would be more comprehensive, although the computation time added may detract from the high-throughput nature of the xTB-ML process.

As discussed previously, the calibration methodology can be expanded beyond TD-DFT. This work additionally applied it to CC2 calculations, but it can be further applied to experimental values. However, this would be time-consuming due to the requirement of real-world measurements. There have been a few previous studies in calibrating TD-DFT against experimental values,[134,135] as outlined in Section 2.2.2, but these used only small experimental datasets. There is a potential here to apply techniques such as text mining to extract experimental excited state data from published papers, though the differences in reporting may make this difficult.

Another potential future work is using xTB-ML in other applications. For example, Chapter 3 expanded candidate space using the graph-based genetic algorithm (GB-GA) developed by Jensen.[144] Instead of using direct ML for energy predictions, it is possible to use xTB-ML. While this would give more accurate results, it would be

significantly slower as xTB-sTDA takes on the order of tens of seconds rather than seconds for ML. However, if calculations are sufficiently parallelized, this process could result in thousands of high-quality candidates being generated in days.

Although xTB-ML is significantly faster than TD-DFT based methods, with comparable accuracy, it is still too slow to screen millions of molecules. As stated previously, xTB-ML can calculate excited states of approximately 1500 molecules per hour (parallelized over 4 nodes), for molecules with <50 heavy atoms such as those in PCQC. Therefore it would take over 3 months to calculate all 3.5M molecules in PCQC (from scratch, starting with only the SMILES string). While a definite improvement over TD-DFT (41 months, using the ground state structure provided in PCQC), this is still slow. Expanding to larger databases with bigger molecules would increase runtime even further. Therefore, an optimized workflow must be developed. The next chapter discusses using active learning to intelligently sample chemical space, intentionally searching for molecules with certain desired properties.

Chapter 5

AL with xTB-ML for high-throughput virtual screening of chromophores

5.1 Motivation

In this chapter, we aim to efficiently and accurately identify candidate chromophores from large-scale databases with high-throughput virtual screening (HTVS). As seen in Section 2.1.1, there are limited large databases with excited state energies already calculated, so these energies must be calculated independently. However, high-accuracy methods are too slow for high-throughput screening, so a faster method such as machine learning or xTB-sTDA must be used instead. Direct ML models may have low accuracy, especially trained on small training sets. xTB-sTDA has also been shown to have low accuracy, but the calibration ML model xTB-ML introduced in Chapter 4 has high accuracy, even when trained on smaller datasets. As mentioned, however, HTVS on large databases (>1M molecules) is time-consuming even with xTB-ML. Instead, we must use active learning (AL) to efficiently sample the chemical space and suggest suitable molecules.

AL was already introduced in Chapter 3, and a similar workflow is presented here, with some key differences. First, xTB-ML is used for labeling molecules instead of TD-DFT. Second, the acquisition function is changed to include not only uncertainty but also suitability. This ensures desired molecules are directly suggested instead of having to run the generated ML model over the entire database at the end. Third, to simplify the workflow, the number of molecules suggested is fixed at 20k, instead of varying the additions based on a fixed uncertainty threshold. The following sections will provide more details about the methods, including dataset descriptions and an overview of the workflow, as well as some results.

Note here that this part of the thesis was done in collaboration with Jiali Li at the National University of Singapore. We developed the specifics of the AL workflow together, but split the computational load to increase efficiency. Jiali ran the ML portion of the workflow as he had access to higher-quality GPUs, while I ran the xTB-ML portion of the workflow since this was more CPU-intensive.

5.2 Methodology

5.2.1 Dataset descriptions

The large-scale molecular database used in this study is PubChemQC (PCQC).[65] While the entire PubChem database could have been used, PCQC is a convenient subsample of PubChem. Details of PCQC are presented in 2.1.1 and 4.2.2. Plots of some properties of PCQC are reproduced in Figure 5.1 for convenience.

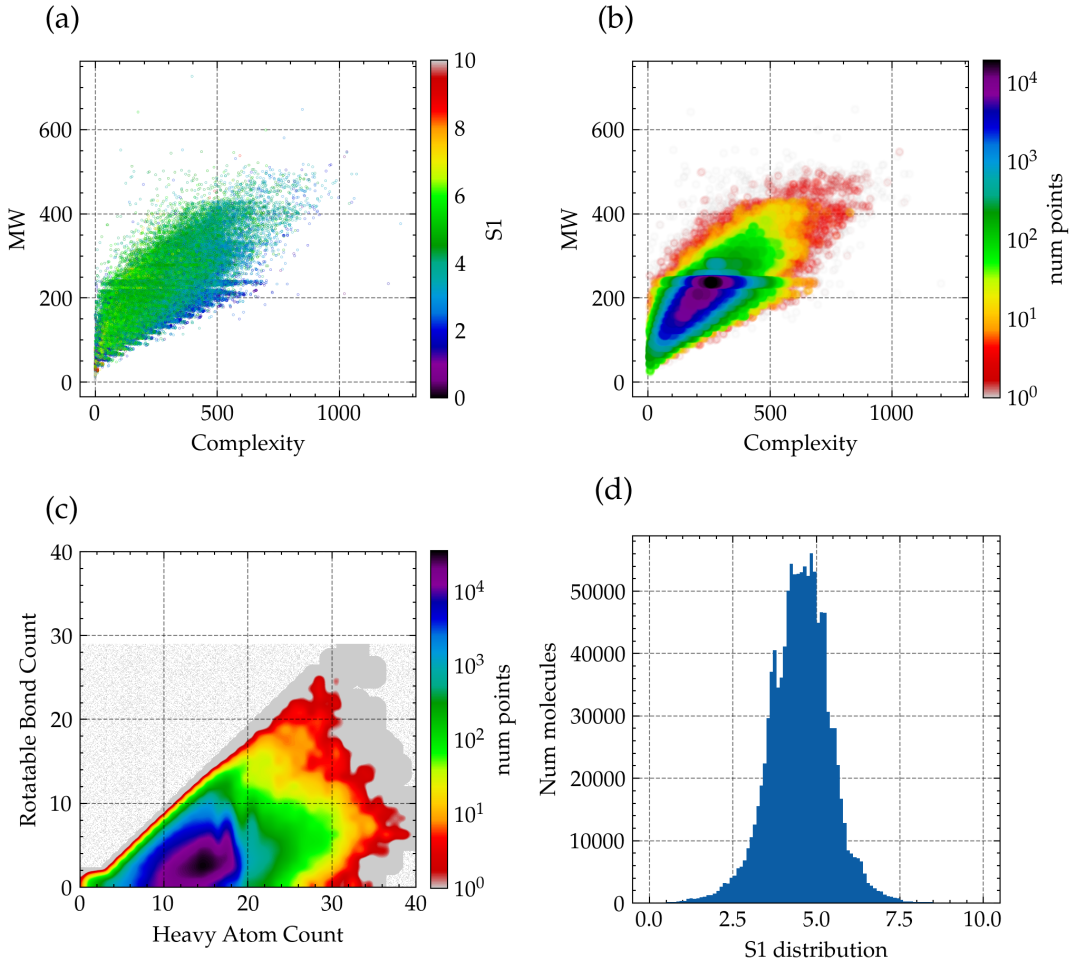


FIGURE 5.1: Plots of properties of molecules in the PCQC dataset. (a) Scatter plot of molecular weight vs. complexity, colored by S1 energy. (b) Density plot of molecular weight vs. complexity. (c) Density plot of rotatable bond count vs. heavy atom ($Z \geq 2$) count. (d) Histogram of S1 energies of all molecules.

As with all AL workflows, an initial training set is required. For this, the SCOP-PCQC dataset presented in Section 4.2.2 was used (more details about the dataset can be found in that chapter). To flesh out the dataset, 5k molecules were randomly selected from PCQC, so the size of the initial training set was 15k molecules.

A fixed test set was chosen to standardize performance evaluation of the AL model at each cycle. 30k molecules were randomly sampled from PCQC, and xTB-ML was used to label each molecule with its S1/T1 excited state data. At each cycle,

the AL model was used to predict excited state properties of the fixed test set, and the accuracy was measured through mean absolute error (MAE).

5.2.2 AL workflow

The purpose of the AL workflow is twofold: to improve the accuracy of the ML model by intelligently expanding the training set, and to suggest suitable chromophores (sensitizers or emitters) by efficiently screening the large database.

Training set expansion is conducted by measuring the uncertainty of the ML model for each molecule in the database. The ML model itself is a 50-ensemble model, with each sub-model having the same architecture (details provided in Appendix Section A.3) but being initialized with different random weights. The uncertainty of each molecule is given by the variance in predictions of the sub-models.

Suitable molecule suggestion is done by averaging the ensemble's predictions of the excited state energy levels and using the energies to calculate a suitability FOM:

$$\varepsilon_{sens} = e^{-A\left(1-\frac{E_{T1}}{E_{S1}}\right)} \tag{5.1}$$

$$\varepsilon_{emit} = e^{-A\left(2-\frac{E_{S1}}{E_{T1}}\right)} \tag{5.2}$$

$$\sigma_{T1} = \frac{\sum_{i=1}^{50}\left(T1_i - T1_{mean}\right)^2}{50} \tag{5.3}$$

$$\sigma_{S1} = \frac{\sum_{i=1}^{50}\left(S1_i - S1_{mean}\right)^2}{50} \tag{5.4}$$

where "sens" refers to sensitizers and "emit" refers to emitters: for sensitizers, T1 should be close to S1, while for emitters, S1 should be around twice T1. The exponential term was used to normalize the suitability from 0 to 1, for mathematical convenience. A is a tunable parameter to adjust the sensitivity of the energy ratio constraint. Note that no strict bounds are defined here, so in Equation 5.1 T1 could be greater than S1, although these molecules are rare and this would likely instead indicate an error in computation. Similarly, in Equation 5.2, if S1 < 2T1, the molecule would be a TTA emitter candidate, while if S1 > 2T1, it would be a singlet fission candidate. Having such flexible suitability functions therefore allows simultaneous suggestion of various types of molecules.

Note further that additional considerations such as strong oscillator strength and spin-orbit coupling are also important characteristics of good sensitizers and emitters, but this study is focused on optimizing energy level alignment instead of reducing efficiency losses. The HTVS approach presented in this work should generate a pool of candidate molecules on which further simulations can then be conducted.

For each AL cycle, the workflow acquires molecules based on an acquisition function defined as the sum of suitability and uncertainty:

$$\begin{aligned} \alpha_{sens} &= B \cdot \varepsilon_{sens} + \sigma_{T1} + \sigma_{S1} \\ \alpha_{emit} &= B \cdot \varepsilon_{emit} + \sigma_{T1} + \sigma_{S1} \end{aligned} \tag{5.5}$$

Where α_i is the acquisition score, σ_i is the uncertainty for either T1 or S1, and B is another tunable parameter, this time to weight the suitability function against the uncertainty and make it more (or less) important, depending on the objective. This definition of the acquisition function ensures the ML model suggests molecules of interest, but also improves in accuracy for all molecules.

The overall AL workflow is presented in Figure 5.2. The initial training data is composed of the 15k molecules discussed in Section 5.2.1. The training dataset is split into 95% actual training and 5% validation for each training epoch. A 50-ensemble ML model is trained on the training data, and used to predict properties of all non-training data, including the fixed test set. Acqusition scores are calculated for the non-training data, as the sum of uncertainty (model variance) and suitability (as discussed above). The 10k molecules with the highest sensitizer acquisition score are chosen, and then the 10k unique molecules with the highest emitter acquisition score are chosen. Note that we specify "unique" emitters because sometimes the uncertainty may dominate the acquisition score, causing overlap between the top 10k sensitizers and emitters, so overlapping molecules are ignored and 10k unique emitters are chosen. Further note that the training set is pruned to avoid overlap with the fixed test set.

Next, on the test set, the MAE is calculated, and on the 20k suggested molecules, the mean suitability scores are calculated. If the MAE is high or the suitability scores are low, the AL cycles continue – the excited state energies of the 20k chosen molecules are labeled with xTB and added to the training set. However, if both the MAE is low and the suitability scores are high, then the model has converged and the AL cycles are completed.

Note here that the xTB-ML calibration model is fixed and not updated with each AL cycle. This is because this would require TDDFT labeling of data, which would detract from the high-throughput nature of this workflow. Also, updating the "ground truth" labeling technique may cause issues with stability. As presented in Chapter 4, the xTB-ML model should be accurate already.

The final results of the AL workflow are all of the suggested molecules compiled from each AL cycle, as well as an optimized ML model that can be used to rapidly screen all remaining molecules in the database to identify further candidates.

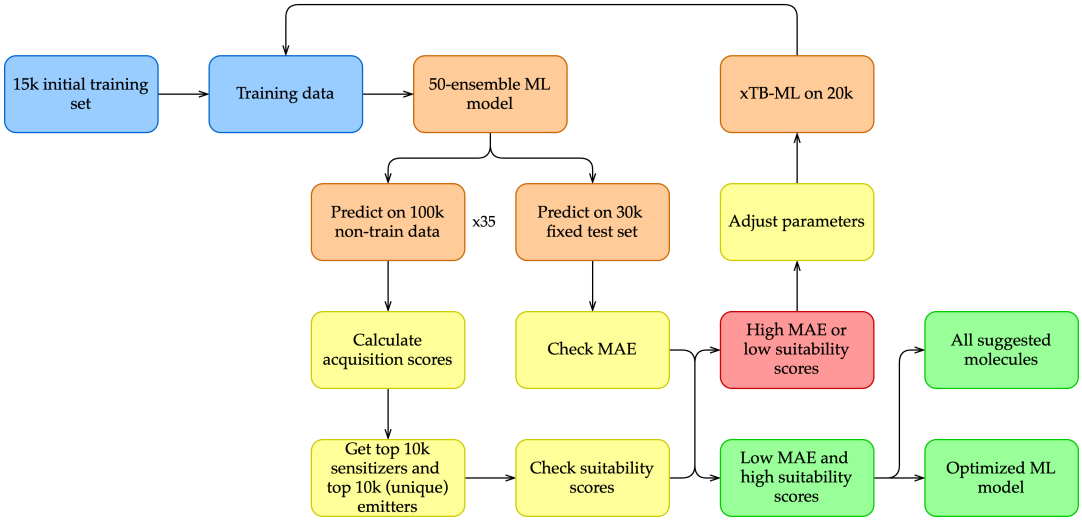


FIGURE 5.2: Overall AL workflow. 50-emsemble ML model is trained on training data, then used to predict properties of all non-training data and the fixed training set. Acquisition scores are calculated for the non-training data, and the top 20k sensitizers and emitters are chosen. The suitability scores of the sensitizers and emitters and the MAE of the test set is calculated. If both properties have not converged, the AL parameters are tuned, the 20k molecules are labeled, and the cycle starts again. If the properties are converged, the optimized ML model is returned. Note that due to memory issues, predictions were made on the 3.5M non-training data in ~35 batches of 100k.

5.3 Results

5.3.1 AL performance

There are two metrics to evaluate the performance of the AL-ML model. First is MAE of the fixed 30k test set, and second is the mean FOM of the suggested 10k sensitizer and 10k emitter molecules. Figure 5.3 shows plots of these two metrics as a function of AL cycle. As seen, the MAE of the fixed test set decreases as the AL cycles progress, but seems to stagnate above 0.2 eV. The mean FOM of suggested sensitizers rapidly increases, but the mean FOM of emitters stagnates around 0.6.

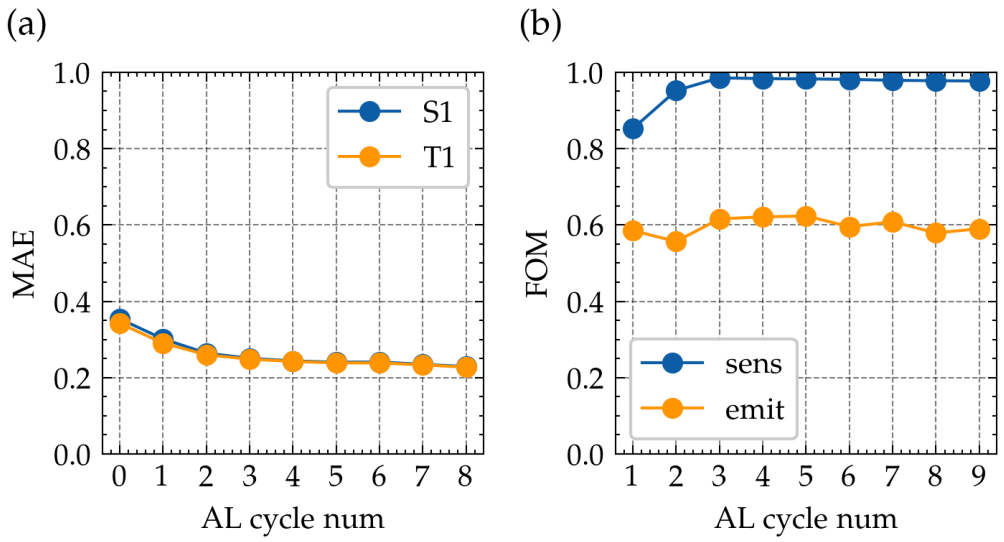


FIGURE 5.3: Plots of improvement of ML model for each AL cycle. (a) Plot of MAE of fixed test set vs. AL cycle, and (b) plot of mean FOM of suggested sensitizer and emitter molecules vs. AL cycle. Note here the numbering system used for AL cycles. Cycle 0 is the initial training set. Each cycle starting with 1 uses AL to add molecules to the training set. Then the total training set so far is used to evaluate the MAE for that cycle. Therefore, cycle 0 has no FOM as molecule suggestions only start with cycle 1.

In the final AL cycle, out of the 10k sensitizers suggested, 9935 had FOM greater than 0.95 (based on Equation 5.1), but out of the 10k emitters suggested, only 767 had FOM greater than 0.95 (based on Equation 5.2). Applying the strict bounds of T1 < S1 required for TTA sensitizers reduces the number to 9916 sensitizers. The number of sensitizers only decreases slightly because normally molecules should not have a higher T1 than S1; if so, that would likely indicate an error in computation. Separating the emitters into TTA emitters (S1 < 2T1) and SF emitters (S1 > 2T1) gives 389 TTA and 378 SF emitters. As expected, because there are no physical constraints on S1 with respect to 2T1, the split between TTA vs. SF emitters is roughly half.

To explore why the mean FOM of the emitters stays so much lower than that of sensitizers, we plot some additional metrics in Figure 5.4. Figure 5.4(a) shows the

number of suggested molecules with high predicted FOM (>0.9), i.e. FOM calculated with predicted S1 and T1 using ML. As seen, the number of suitable sensitizers rises quickly to 10k, while the number of suitable emitters initially rises, but then falls rapidly. A similar trend is seen in Figure 5.4(b), where the number of high-uncertainty molecules (>0.5, defined as the sum of uncertainty terms in the acquisition function over the overall acquisition score) is plotted against AL cycles. The number of high-uncertainty sensitizers drops rapidly to essentially 0, while the emitters initially see a rapid drop, but then a more gradual increase. These trends seem to suggest that the AL model has exhausted its search for emitters after 4 AL cycles. Instead, after 4 cycles, the model seems to be expanding its reach by including more high-uncertainty molecules.

The suggested molecules were then run with xTB-ML to confirm results, and the number of molecules with high confirmed FOM (>0.9) are shown in Figure 5.4(c). A similar trend to Figure 5.4(a) is seen, but at a more modest scale. Almost all sensitizers again exhibit high FOM, while the emitters increase until cycle 5 and then decline. In contrast to Figure 5.4(a), however, the increase and decrease are much more gradual, and the overall number of confirmed emitters hovers around 1500 per cycle.

To investigate why the number of confirmed emitters does not increase, we plot accuracy calculated two ways in Figure 5.4(d). First, the simplest calculation is the number of confirmed high-FOM molecules over suggested high-FOM molecules, shown as "sens-1" and "emit-1." Sensitizer accuracy approaches unity, and at first glance it appears emitter accuracy similarly explodes after cycle 4. However, this is not a true measure of accuracy. Instead, "sens-2" and "emit-2" show the fraction of suggested high-FOM molecules that were confirmed to have high FOM. This is distinct from the former definition as the number of confirmed high-FOM molecules may include molecules that were not suggested to have high FOM but ended up exhibiting these properties anyway. In reality, the accuracy of the suggested molecules remains low, at around 17%.

This is a peculiar result, as it is unclear why molecules suggested to have low FOM would instead be confirmed to high a FOM. To explore this further, we can plot MAE per energy interval, as shown in Figure 5.5. As seen, initially the errors are large, with many MAEs higher than the energy interval itself. However, as the cycles progress, the MAEs decrease significantly. Generally, the shape of the MAE vs. energy plot matches the shape of the histogram of energies, such that energy intervals with more molecules have lower MAE. The high MAEs for certain energy intervals could help explain why so many molecules are unexpectedly showing high FOM, as they would initially not be identified as suitable with the AL-ML surrogate model.

It is now useful to get a more visual representation of the molecular space to qualitatively explore some of the phenomena discussed above.

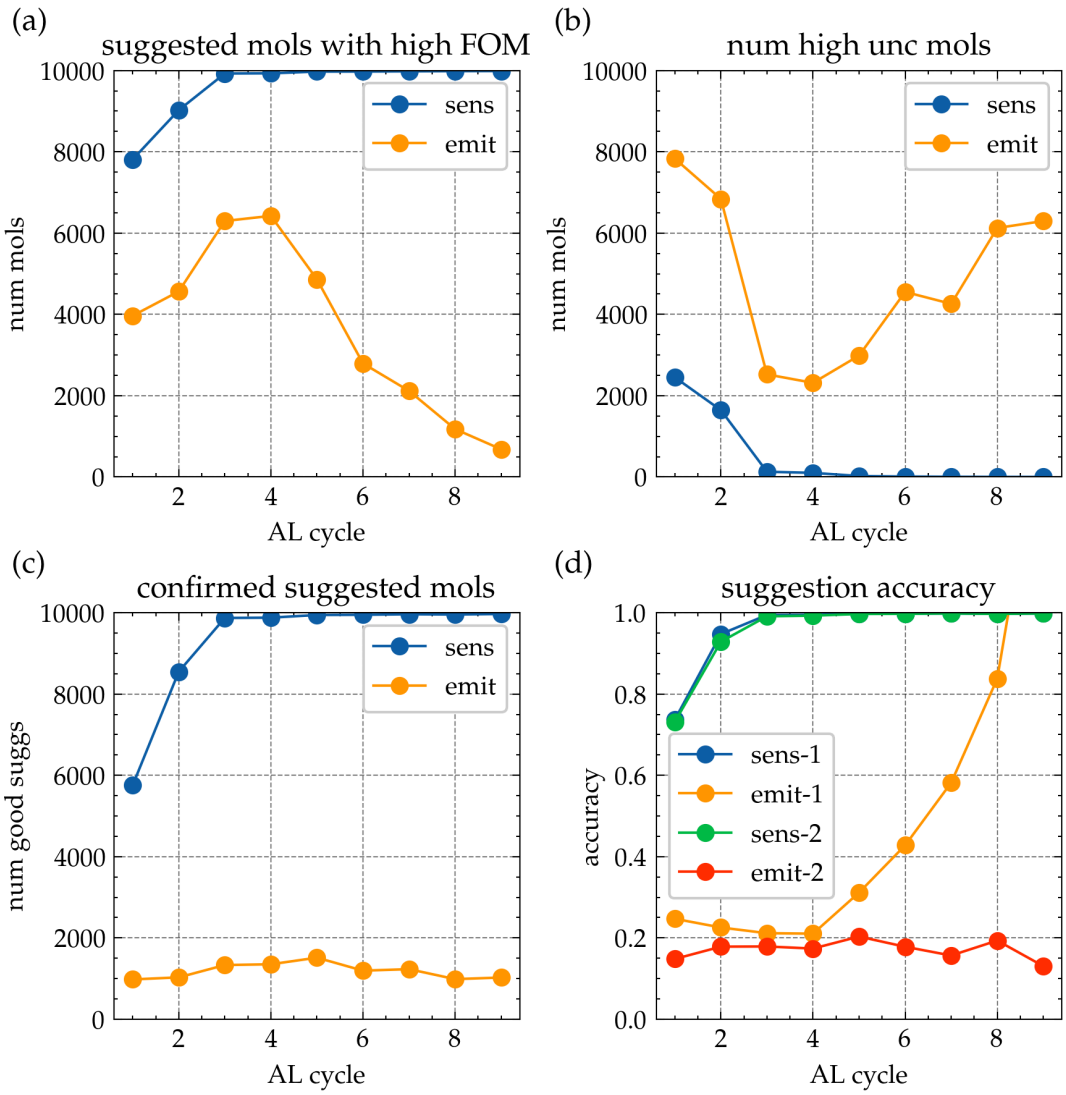


FIGURE 5.4: Additional metrics for AL cycles. (a) Number of molecules with high FOM (>0.9) as determined by surrogate ML predictions. (b) Number of molecules with high uncertainty fraction (>0.5), defined as sum of both uncertainty values over the acquisition score. (c) Number of suggested molecules confirmed to be suitable with xTB-ML. (d) Suggestion accuracy, (1) all confirmed molecules over predicted high-FOM molecules, and (2) fraction of predicted high-FOM confirmed with xTB-ML

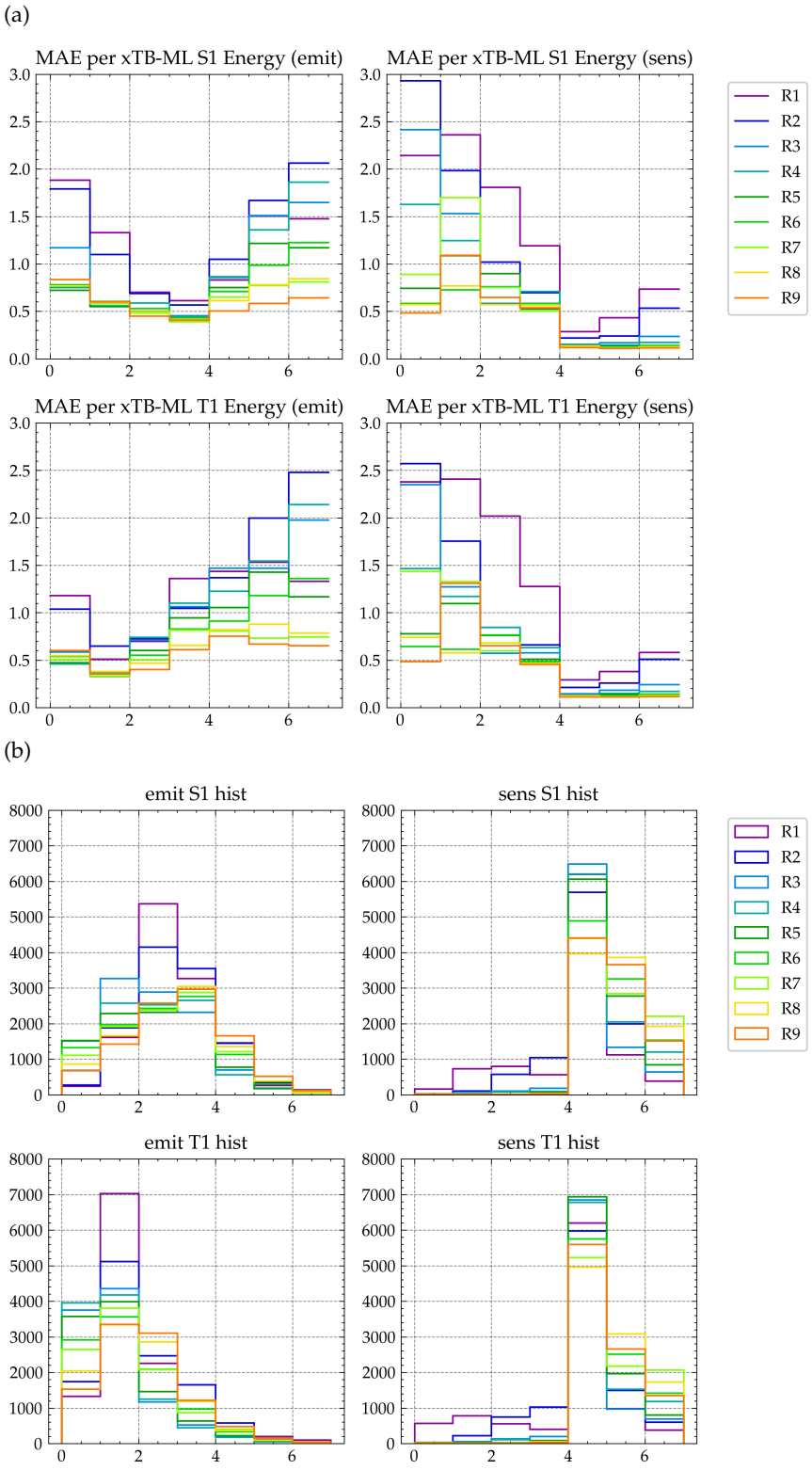


FIGURE 5.5: (a) Histograms of MAE per energy interval of xTB-ML, for S1/T1 energies in sensitizers and emitters. (b) histograms of S1 and T1 energies of emitters and sensitizers. Note that these MAEs are different from the MAEs calculated in Figure 5.3, as this figure uses the 20k suggested molecules as the "test set" while the former uses a fixed 30k test set.

5.3.2 Chemical space mapping

Figure 5.6(a) shows where the molecules added at each AL cycle are located in global chemical space. As seen, while the molecules are initially distributed widely, by cycle 4 they start to cluster in certain regions. By the last cycle, there are definitive clusters in a few regions of chemical space, but there is simultaneously broad coverage of chemical space. Further, Figure 5.6(b) shows the distribution of sensitizers (red) and emitters (blue) separately. As seen, there are clear boundaries between the two types of molecules, for example the top left being almost all emitters, while the right side is majority sensitizers. It further looks as though sensitizers are clustered, implying a region has been found with high sensitizer potential, while emitters are still spread out, so an ideal region has not been found yet. To explore this more deeply, Figure 5.6(c) shows the uncertainty/acquisition score fraction, i.e. if the molecule was included due to high uncertainty or high suitability. As seen, most sensitizers have nearly zero uncertainty, indicating they were chosen due to their properties. In contrast, the emitters generally have higher uncertainty, indicating emitters with ideal properties have not been found yet. This is not always the case, however – for example, the tip on the top left includes emitters with low uncertainty ratio, indicating that may be a promising region.

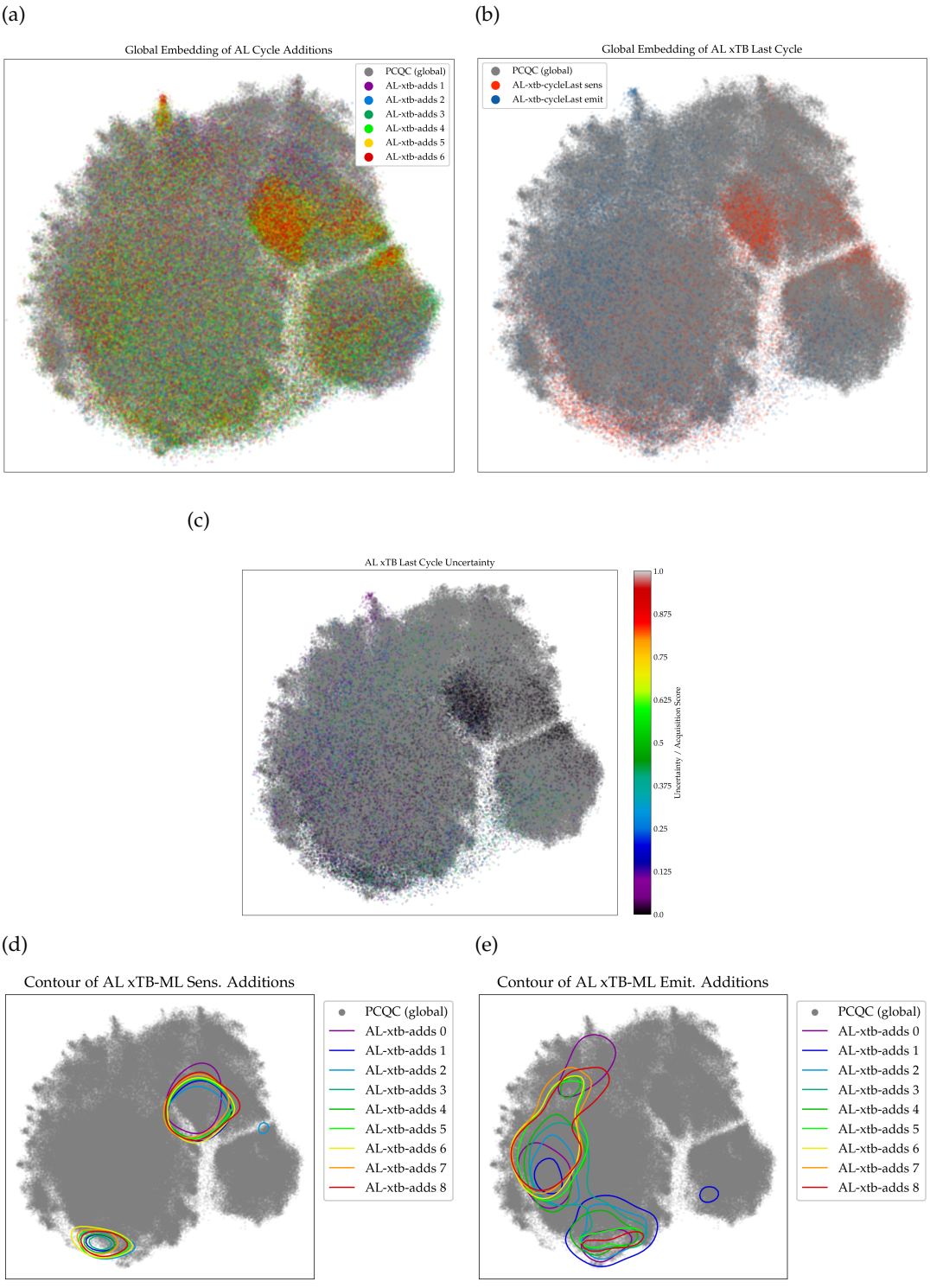


FIGURE 5.6: Global chemical space embedding of various datasets. (a) Locations of molecules added at each AL cycle, compared to global chemical space. (b) Locations of sensitizers (red) and emitters (blue) separately plotted for the last AL cycle. (c) All last-cycle AL molecules colored by uncertainty/acquisition score ratio. Lower ratio means uncertainty contributed less to the score than suitability, higher is vice versa. Global embedding generated by 350k randomly sampled PCQC molecules, using UMAP based on the 2D Jaccard similarity between Morgan fingerprints. Embedding of AL molecules predicted based on global embedding.

5.3.3 Identifying chromophores

In all AL cycles, there were a total of 88056 unique sensitizers and 79860 unique emitters that were suggested, and of those, 79149 TTA sensitizers, 5781 TTA emitters, and 4222 SF emitters were confirmed to be suitable by xTB-ML, using the strict bounds defined in Equation 3.5. Recall that in Chapter 3, 307216 sensitizers, 2763 TTA emitters, and 1694 SF emitters were identified. While not all of them were confirmed with TD-DFT, these numbers can serve as order-of-magnitude approximations to the number of potential candidates. This suggests that AL is likely close to saturation for emitters, but further AL cycles should continue to idenfity sensitizers.

To get a sense of the types of molecules being identified as sensitizers or emitters, it is useful to visualize the scaffolds of these molecules. RDKit's MurckoScaffold module was used for this purpose. Figure 5.7 shows the 32 most common scaffolds in identified suitable (a) sensitizers and (b) emitters (without differentiating between TTA or SF emitters), where each scaffold must have at least 10 heavy atoms.

As seen, there are a few stark differences between suggested sensitizers and emitters. First, there are no aromatic rings in the sensitizers, while the emitters have several. The emitters have majority 6-carbon rings, while many of the sensitizers have 3- or 5- carbon rings. Many of the emitters have oxygen atoms, while few of the sensitizers do. There are also some similarities, such as many scaffolds of both sensitizers and emitters are composed of rings connected with a bond or a chain.

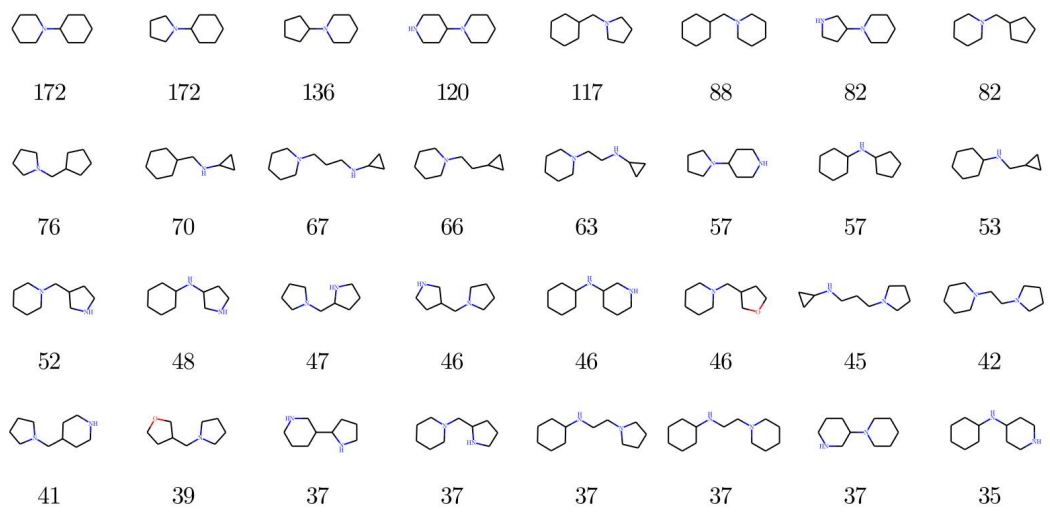
To better understand the relationships between identified molecules, it is possible to generate a graph representation of the dataset. The graph features molecules represented by nodes, connected with edges if their similarity score is greater than 0.5. After computing these properties, the graph representation can be visualized with Argo Lite,[159] as shown in Figure 5.8 below.

This graph representation is useful to evaluate the quality of molecules within the database. Molecules with high degree or high page rank would be more likely to be suitable emitters. Degree is a measure of connectivity and counts the number of connections to other nodes. Page rank is a measure of importance, including the number of connections but also the importance of those connections. While the graph presented here only includes molecules labeled with xTB-ML, it can also be used as a measure of confidence for ML-predicted molecules – if a molecule has a high page rank when added to the graph, it is likely to be suitable.

The 8 molecules with the highest degree and page rank are shown in Figure 5.9. The SMILES, S1, T1, S1/T1 ratio, and degree/pagerank for these molecules is available on GitHub.[142]

The AL workflow will give all potential TTA molecules based on ratios of S1/T1. To identify specific molecules for our energy region of interest, we need to screen the molecules for their energy levels and match sensitizers to emitters to identify pairs for TTA. For NIR-TTA, we want sensitizer S1 to be around 1.1 eV (1.0 to 1.2 eV) and emitter S1 to be around 2.2 eV (2.0 to 2.4 eV).

(a)



(b)

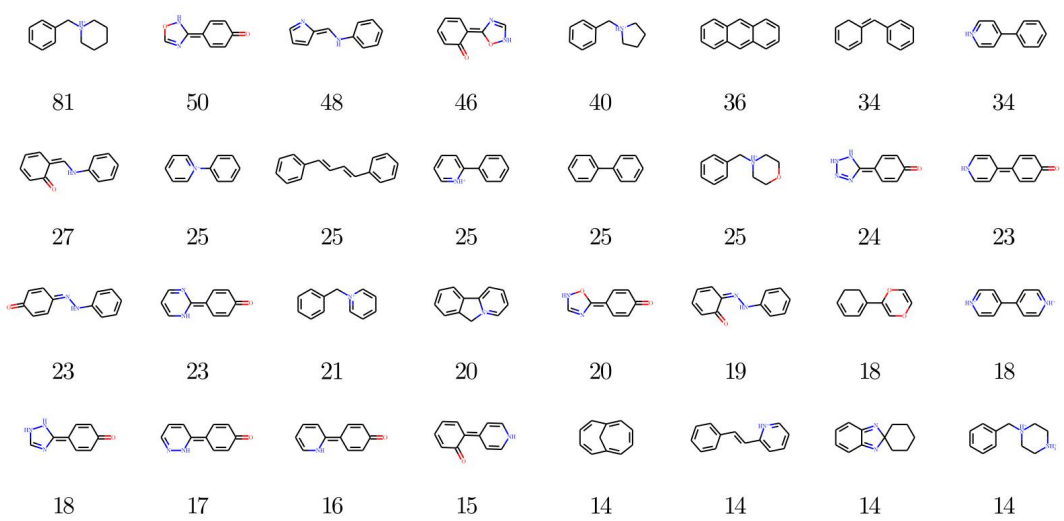


FIGURE 5.7: Most common scaffolds for suitable (a) sensitizers and (b) emitters identified by AL-xTB-ML. Each scaffold must have at least 10 heavy atoms (Z>1). Numbers below each scaffold shows number of times it appears in the dataset of suitable molecules.

Applying these constraints to the identified sensitizers and emitters gave 15 potential sensitizers and 322 potential emitters. The reason there are so fewer sensitizers than emitters is due to the distribution of energies, as seen in the histogram in Figure 5.5(b). The sensitizer S1 distribution is skewed such that most S1 energies are above 4 eV. In contrast, the emitter S1 distribution is more normal, with a peak around 3-4 eV but still a substantial number of molecules around 2 eV. Molecular information (including SMILES, S1, T1, and FOM) for all identified chromophores, as well as NIR identified chromophores, is available on GitHub.[142]

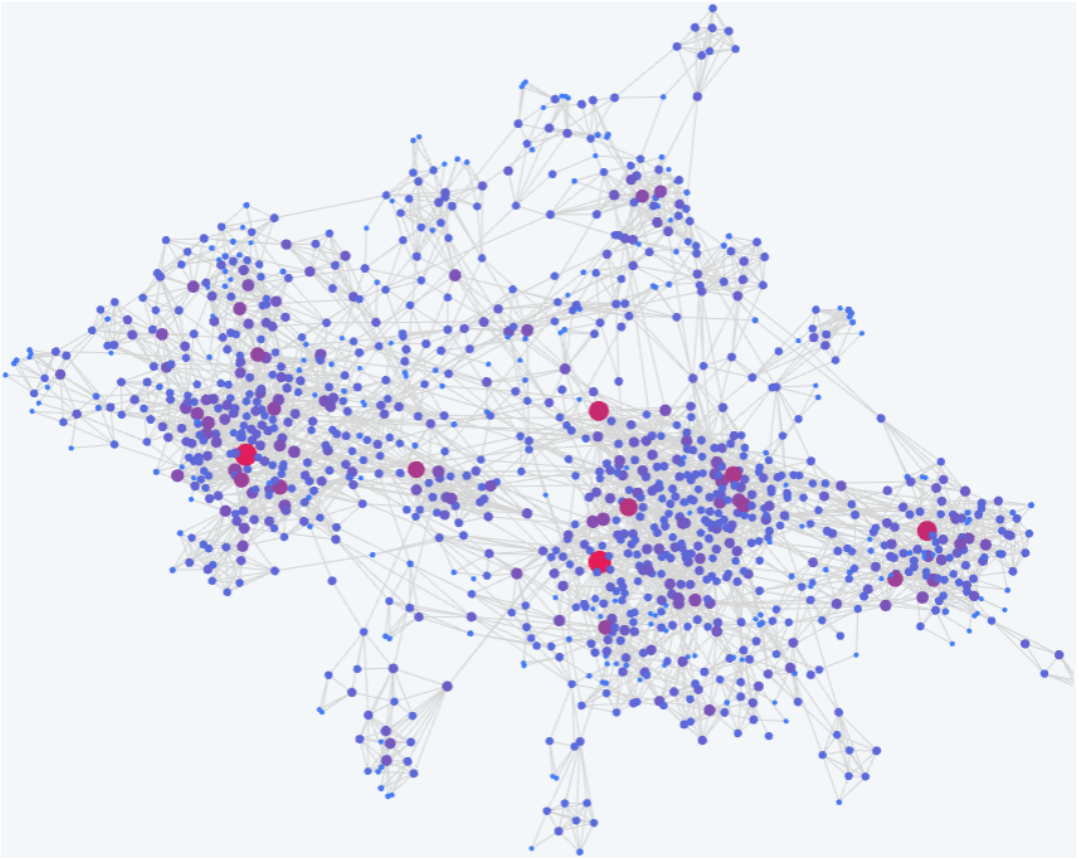
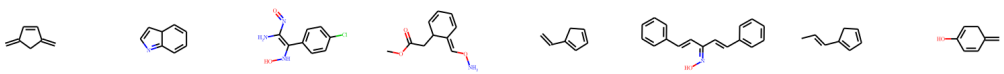


FIGURE 5.8: Argo Lite graph representation of identified emitters. Each node represents a molecule in the dataset, while an edge is made if the two connected molecules have similarity scores greater than 0.5. Size and color of each node correspond to the degree of connection. Graph representation generated by Jiali Li.

(a)

32 31 27 27 24 22 22 20

(b)



6.37 4.83 4.10 4.08 3.99 3.71 3.68 3.33

FIGURE 5.9: Highest ranked molecules by (a) degree and (b) page rank in graph representation of emitter dataset. Legend indicates score of each molecule.

5.3.4 Improvements to AL workflow

There is clearly room for improvement to the AL workflow above, especially for emitter identification. We propose a few changes to the acquisition function of emitters to help improve efficiency and accuracy of selection. Instead of having only two terms, one for suitability and one for uncertainty, it may be beneficial to include a similarity term. This term would include comparison of a test molecule to all previously identified emitters, as such molecules are likely to have similarly suitable properties. This would focus the search in certain areas of chemical space where emitters are likely to reside. The existing acquisition function seems to work until cycle 5, at which point the acquisition function would be amended to include the similarity term, i.e.

$$\begin{aligned} \alpha_{sens} &= 2 \cdot \varepsilon_{sens} + \sigma_{T1} + \sigma_{S1} + 4 \cdot \text{Sim} \\ \alpha_{emit} &= 2 \cdot \varepsilon_{emit} + \sigma_{T1} + \sigma_{S1} + 4 \cdot \text{Sim} \end{aligned} \tag{5.6}$$

where ε, σ_{T1}, and σ_{S1} are normalized between 0 and 1, and Sim is the similarity score.

We have yet to comment on the form of the similarity score. Fundamentally, we would use cosine similarity between Morgan fingerprints of molecules. However, the algorithm of calculating similarity of test molecules to identified molecules must be carefully designed. Taking the maximum of all similarity scores may result in spurious matches if the test molecule is similar to only 1 or 2 existing molecules. On the other hand, taking the average of all similarity scores would flatten the data significantly. Instead, a cluster similarity method is proposed. First, the pre-identified molecules are clustered in chemical space. Then, for each cluster, the similarity score of the test molecule to all molecules in the cluster is calculated, and the average of the top 100 similarity scores is used. The maximum average similarity score across the clusters is then used as the final similarity term in the acquisition function. This ensures no spurious molecules are included, and only molecules definitively matching existing clusters are added.

Another possible addition to the acquisition function would be to even out the distribution of molecules at different energy intervals. Currently, as seen in Figure 5.5, the MAE is not uniform across all energy levels, but seems to improve when more molecules are located in that energy interval. To force the model to learn more about energies with high error, a term could be added in the acquisition function to make molecules with these energies more likely to be chosen. This would help improve the overall MAE of the model, and prevent it from stagnating as it seems to be from Figure 5.3.

These new acquisition functions are being tested at the time of writing, so results are unfortunately not available yet. The new acquisition functions will be compared to the original acquisition function, as well as a control acquisition function with no suitability term.

5.4 Conclusions and Future Work

This chapter applies xTB-ML to screen large-scale databases, using active learning to intelligently sample molecules of interest. The two main differences between the active learning workflow implemented here and that of Chapter 3 is the inclusion of suitability in the acquisition function, and the use of xTB-ML as the labeling technique instead of TD-DFT. Adding suitability allows active suggestion of candidate molecules in each AL cycle. The uncertainty term is still included in acquisition, so the model's accuracy should also improve. Using xTB-ML instead of TD-DFT allows rapid iterations of AL. For example, each AL xTB-ML cycle takes approximately 2 days – 24 hours for the surrogate ML training and prediction and 15 hours for xTB-ML labeling of the top 20k suggested sensitizers and emitters. In contrast, each AL TDDFT cycle takes approximately 2 weeks, with the bulk of that consumed by TDDFT labeling. A natural concern with replacing TDDFT with xTB-ML is accuracy. However, as shown in Chapter 4, xTB-ML should predict TDDFT values within 0.15 eV. While this may vary for specific molecules, this is a good point of reference.

The main result from this work is the rapid nature of the AL cycles. This allows multiple workflows to be quickly tested and evaluated. In this chapter we have only presented one AL workflow based on an acquisition function including suitability and uncertainty; however, multiple additional workflows are proposed, and many others can also be tested. If desired, the best workflow can then be used on an AL methodology using TDDFT, for more accurate molecular suggestions.

This chapter focuses on developing the methodology for rapid AL, but evaluating the efficacy of the AL workflow is limited to calculation with xTB-ML. More intensive evaluation is delegated to future work. The easiest way to evaluate a workflow would be to compare the results to molecules output in Chapter 3. However, the accuracy of the previous workflow was low, as it used direct ML predictions, with only a small fraction of suggested molecules were confirmed with TDDFT. The best way of evaluating the AL workflow would be to confirm high-scoring molecules with TDDFT. Despite the computational expense of TDDFT, if only the top ~1000 or so molecules are calculated, this could be a reasonable evaluation technique.

Another avenue of immediate future work is to expand the number of candidates. This can be done by running the final optimized ML model on all 3.5M molecules in PCQC. This should give several additional potential sensitizers and emitters, which can be either confirmed with xTB-ML, or added to the graph representation to evaluate confidence in the predictions. Another option for expansion is using the graph-based genetic algorithm (GBGA) presented in Section 3.3.4. Since an output of this AL workflow is an optimized ML model for prediction of S1 and T1 energies, this can directly replace the ML model implemented in the GBGA workflow in Chapter 3.

Finally, the AL workflow can be applied to high-throughput virtual screening of other datasets. Since no TD-DFT is involved, and xTB-ML is fast, datasets can be

rapidly screened and potential candidates suggested. The best AL workflow, as still to be determined, can be generalized to other databases, as the acquisition function is not unique to PCQC. For a time-scale reference, the 3.5M molecule PCQC database was screened in 2 weeks. The time requirement would fortunately scale nicely, as the xTB-ML model will always be labeling the top 20k molecules, so this would be constant, and the ML model scales with the number of features and the number of neurons per layer, which is faster than linearly. It would therefore be possible to screen tens or hundreds of millions of molecules quickly using an optimized AL workflow.

Overall, this chapter combines the two techniques presented in Chapters 3 and 4, active learning and xTB-ML, to intelligently sample a large molecular database and actively suggest candidates. While the workflow still requires optimization, this is a useful starting point. The following chapter will summarize the entire thesis and present some final thoughts of the work.

Chapter 6

Conclusion

6.1 Summary

In this thesis, we have used high-throughput virtual screening (HTVS) of large molecular databases to identify potential chromophores for triplet-triplet annihilation (TTA) and singlet fission (SF). As outlined in Chapter 1, TTA and SF materials can be used to shape the solar spectrum to be more suitable for existing solar cells, increasing their maximum theoretical efficiency from 33% to around 50%. Unfortunately, photon conversion materials used in TTA and SF suffer from a variety of losses, categorized under either efficiency losses, which reduce the probability of an absorbed photon being re-emitted, or energy losses, reducing the output energy of the re-emitted photon. We have focused on reducing energy losses by discovering molecules with optimized energy level alignment. For example, for SF, the singlet energy (S1) should be just above twice the triplet energy (T1). TTA requires two molecules, sensitizers and emitters, where sensitizers have S1 just above T1 and emitters have S1 just below twice T1. The HTVS process involves calculating the S1 and T1 energies for molecules and selecting the molecules with the most optimal energy level alignment.

There are various ways of calculating the S1 and T1 excited state energies. The most accurate would be do use post-Hartree Fock *ab initio* methods, including coupled cluster calculations. Unfortunately, these are incredibly time intensive. The most common excited state method is time-dependent density functional theory (TD-DFT), which has a reasonable tradeoff between computational cost and accuracy, especially with the development of recent functionals. However, for HTVS, even TD-DFT is too slow. Recently, several high-throughput computational techniques have been developed, including extended tight binding (xTB) methods which can be combined with the simplified Tamm-Dancoff approximation (sTDA) for ultrafast computation of excited state energies, on the order of seconds to minutes.

The above techniques attempt to find solutions to Schrödinger's equation, with varying levels of approximations, and are therefore classified as computational chemistry techniques. However, it is also possible to apply a data-driven approach for excited state energies. Machine learning (ML) models such as neural networks can be trained on large datasets and used to rapidly predict excited state energies of

thousands of molecules within seconds. Training set generation can be done either randomly or with active learning (AL), which more purposefully forms the training set with high-uncertainty molecules. Chapter 2 provides more details about the methodologies presented above.

The results chapters of this thesis, Chapters 3 - 5, used various combinations of the above techniques to conduct HTVS to identify potential chromophores.

The first approach, as presented in Chapter 3, was to screen existing high-accuracy databases (with TD-DFT level calculations) for potential chromophores. Unfortunately, a large-scale triplet energy database does not exist. The large-scale quantum chemistry database PubChemQC (PCQC) contains singlet energy TDDFT calculations for 3.5M molecules, but calculating 3.5M triplet energies would be prohibitively expensive. Instead, our approach was to use ML to predict energies based on a smaller (<10% the size of PCQC) training set. Training set generation was done sequentially with AL, by training a model, using it to predict uncertainty in the remaining molecules, selecting high-uncertainty molecules, running TD-DFT on those molecules, adding them to the training set, and repeating this cycle. Because singlet energies were already available in the database, we could rapidly conduct AL cycles, and after 8 cycles we achieved an MAE of 0.16 eV, using a training set of 276k molecules. For triplet energies, time constraints limited the AL cycles to 1, but an MAE of 0.3 eV was still achieved, using a training set of 133k molecules. Using these 2 ML models, we were able to rapidly screen all 3.5M molecules in PCQC, identifying 307,216 sensitizers, 2763 TTA emitters, and 1694 SF emitters. Of interest to solar applications are NIR TTA materials, so restricting the energy levels to that region gave ~3000 molecules, and running these with TD-DFT gave 7 confirmed sensitizers and 7 confirmed emitters. To expand the candidate space, the graph-based genetic algorithm (GBGA) was updated to use the 2 ML models generated in this study. Running GB-GA output ~5000 total sensitizers and emitters of interest.

There are unfortunately a few limitations with using ML to directly predict excited state energies, including issues with accuracy, the large training set required, and its inherent black-box nature. Therefore, the second approach, as presented in Chapter 4, uses ML to calibrate a high-throughput computational technique (xTB-sTDA) against TD-DFT, instead of directly predicting excited state energies. The training set for calibration requires both xTB-sTDA and TD-DFT calculated values for molecules. While xTB-sTDA values can be calculated locally, TD-DFT calculations are time-consuming, so these values were taken from databases instead. The two databases considered were SCOP-PCQC, an independently generated subset of PCQC consisting of 10k molecules relevant to TTA/SF, and QM-symex-10k, composed of 10k randomly selected molecules from 173k radially symmetric molecules. Expansions to these datasets were also considered, including the 105k/107k S1/T1 AL datasets from Chapter 3 and the full 173k QM-symex dataset. ML models were trained on these various datasets, taking the molecular SMILES string and the error (TDDFT – xTB-sTDA) as input, and used to predict errors for various blind test

sets. These blind test sets included 1,143 small aromatic molecules from Wilbraham et al.[112] and 1,000 indolonaphthyridine thiophene derivates from Fallon et al..[52] The two ML models that performed the best were the initial 20k training set (SCOP-PCQC + QM-symex-10k), with an average MAE of 0.150 eV, and the 300k training set with all expansions, with an average MAE of 0.161 eV. Both of these vastly improve the average MAE of 0.341 eV for raw xTB-sTDA values. The xTB-ML-20k model was then used for various applications, including identifying 60 NIR-TTA chromophores among the 250k small aromatic molecules in Wilbraham et al.[112] and mapping the accuracy of xTB-sTDA in chemical space. The same methodology used to calibrate xTB-sTDA against TD-DFT was then used to calibrate xTB-sTDA against coupled cluster (CC2) values in QM8, finding the ML-calibred xTB values outperformed TD-DFT values, with an MAE of 0.15 eV compared with 0.26 eV for PBE0 and 0.19 eV for CAM-B3LYP. These results are extremely promising, showing ML can help increase the accuracy of xTB-sTDA and improve its predictive performance, despite it being a high-throughput technique.

Finally, the third approach, as presented in Chapter 5, combines the above two approaches. While calculating 3.5M molecules with xTB-sTDA would be 2-3 orders of magnitude faster than TD-DFT, it would still be relatively slow (3 months, when parallelized over 4 nodes). Therefore, we instead use AL to intelligently sample the 3.5M molecular space for potentially suitable chromophores. The AL workflow is similar to that of Chapter 3, with some critical differences. Because we want to actively suggest potential chromophores, instead of only selecting the high-uncertainty molecules, the workflow also selects molecules with high suitability. Then, the selected molecules are labeled with xTB-ML instead of TD-DFT. Its efficacy as a direct replacement to TD-DFT is confirmed by the analysis in Chapter 4, which demonstrates a low MAE for xTB-ML. By using xTB-ML, this workflow is significantly faster, allowing 9 AL cycles to be completed in under 2 weeks. The final ML model had an MAE of 0.23 eV using a 192k training set. Compiling all suggestions from the 9 cycles gave 79149 TTA sensitizers, 5781 TTA emitters, and 4222 SF emitters confirmed with xTB-ML, of which ~350 were potential NIR-TTA chromophores. Based on performance metrics and chemical space maps of selected data, it seems that the emitter space has been fully explored and candidates output, while more sensitizers could be suggested with further AL cycles. Although the chapter only presents one acquisition function, others may help improve results. For example, in addition to uncertainty and suitability, we could add a similarity term or energy term, to optimize suggested molecules and reduce error.

All 3 of the above approaches to HTVS gave promising results, and should be widely applicable beyond the specific molecular characteristics investigated here. Each results chapter has a short discussion of future work at the end, but a broader outlook is presented here.

6.2 Outlook

Of the 3 results chapters presented in this work, Chapter 4 is perhaps the most applicable to other studies. The xTB family of methods is increasing in popularity due to its computational efficiency, and is often used in HTVS studies. While this thesis calibrated xTB-sTDA excited state data, it is absolutely possible to expand the ML calibration technique to other properties. As shown, the calibration methodology is flexible, as the reference was easily switched from TD-DFT to CC2 values. Other desired properties, such as ground state energy, HOMO/LUMO gap, Fermi-level, vibrational frequencies, thermochemical properties, 3D structure, etc. can also be calibrated, given TD-DFT (or other) reference values. Note that the number of datapoints required may be larger than the training set used in this work, depending on the complexity of the desired property. This idea of ML calibration of xTB is not completely unheard of – in the generation of the xTB methods, an extensive parameterization is required, which uses a fit on a large dataset. While not specified as an ML calibration, a high-order fit is functionally similar. An ML fit may not be as generalizable as the xTB parameterization, but would help increase accuracy of a desired property.

The methodology presented in Chapter 5 is also widely applicable. Most existing active learning studies for chemical exploration use TD-DFT for data labeling. However, this limits the comparison of different AL workflows and techniques. The primary value of the novel proposed workflow is its speed, allowing testing of various acquisition functions. Using a workflow similar to the one proposed in this work, other studies could quickly get a sense of the size of the candidate space, note potential optimizations to the AL workflow (i.e. adding terms to the acquisition function), and get a sense of how many molecules are required for low MAE predictions. Then, with an optimized AL workflow, they could return to a high-accuracy data labeling technique. Having a high-speed AL workflow also allows rapid screening of massive databases, i.e. >10M molecules, that would be impossible to screen with conventional TD-DFT based AL workflows.

The results from Chapter 3 are the most applicable portion of that work. A new triplet energy database is generated with ML, and can be used for additional materials screening purposes beyond NIR-TTA materials, for example UV-to-Vis SF materials. The GB-GA update presented also demonstrates the versatility of the method. The algorithm had several scoring functions already implemented, including partition coefficient and xTB-sTDA based absorption (without xTB ground state optimization). Adding the ML models for S1 and T1 energies shows that the algorithm can work with a variety of scoring functions.

Beyond applying the methods/data generated to other studies, there is also potential for expanding the scope of the study itself. As mentioned in the Introduction, this study only considers vertical excitation energies. However, excited state relaxation may be important for some systems. Adding this phenomenon would allow

vertical emission and adiabatic energy calculations. There are a few ideas to incorporate this in a HTVS workflow. In this study, since xTB-sTDA only calculates vertical excitation, for consistency the TD-DFT workflow was standardized to this setting as well. However, it is possible to have TD-DFT calculate the adiabatic energy (with excited state relaxation) instead, and then calibrate the xTB-sTDA vertical excitation energy against the TD-DFT adiabatic energy. Or, taking it a step further, TD-DFT could calculate the vertical emission using the excited state relaxed structure, and xTB-sTDA vertical excitation could be calibrated against that. Alternatively, a separate ML calibration could be generated for excited state structure relaxation, comparing the xTB-generated structure against the TD-DFT relaxed structure. Then, using the ML-calibrated structure, a single-point xTB-sTDA calculation could be done to get the vertical emission energy. In general, as seen, there are several options for incorporating excited state dynamics into the xTB-ML workflow, which could be the topic of extensive future research. The main question is its compatibility with HTVS, since adding additional TD-DFT calculations would increase computation time.

The other potential expansion of scope is incorporating efficiency loss into the HTVS workflows. As discussed earlier, this study focuses on optimizing energy level alignment to reduce energy loss, ensuring the emitted photon is essentially twice the energy of the two absorbed photons (for TTA). The other major loss mechanism for TTA is efficiency loss, reducing the probability that an absorbed photon will be re-emitted. This is composed of several terms: oscillator strength (OS) from ground to excited state, intersystem crossing (ISC) from singlet to triple states in the sensitizer, triplet-triplet energy transfer (TTET) from sensitizer to emitter, and triplet triplet annihilation (TTA) between two excited emitters. OS is the probability of absorption and is already calculated with sTDA/TDDFT. ISC is a function of spin-orbit coupling (SOC), which is calculable from excited-state singlet and triplet wavefunctions. For example, PySOC[160] calculates SOC using outputs from Gaussian or DFTB+. TTET and TTA are forms of Dexter energy transfer, which is a function of wavefunction and spectral overlap, so again is calculable. The main question is whether these calculations can be included in a HTVS workflow, or whether they are too expensive. Regardless, these calculations should at least be conducted on the molecules suggested in this study, to see if any low energy-loss, high-efficiency molecules exist.

Overall, we hope the combination of high-throughput computational chemistry and machine learning presented in this study will spark further investigation and help improve the accuracy of high-throughput techniques through data-driven approaches. As is evident, various applications exist, and we are excited to see where this work goes.

Appendix A

Supplementary Information

A.1 Conventional ML for AL

The following plot shows a histogram of S1 errors (predicted vs. true) using a randomly sampled 500k training set tested on 350k molecules.

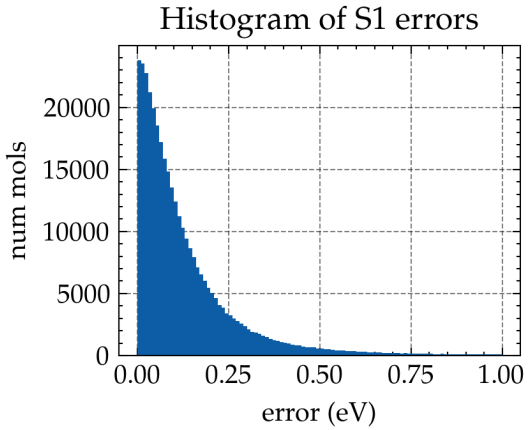


FIGURE A.1: Histogram of S1 errors, for ML model trained on 500k randomly sampled molecules and tested on 350k molecules.

A.2 Molecular data for identified chromophores

A.2.1 Strict NIR bounds

Table A.1 shows the identified chromophores using strict NIR bounds. S1, T1, and FOM refer to the predicted values using the AL-ML model.

A.2.2 Loose NIR bounds

Table A.2 shows the identified chromophores using loose NIR bounds. S1, T1, and FOM refer to the predicted values using the AL-ML model.

TABLE A.1: AL-ML candidates using strict NIR bounds

SMILES	S1	T1	FOM	S1 TDDFT	T1 TDDFT	FOM TDDFT	type
[CH2-]N1CCCC2C1CCCC2	1.121	1.101	0.983	1.156	1.114	0.964	sens
CCC1=NC=C2N1CCN(C2)[CH2-]	1.174	1.142	0.973	1.161	1.156	0.996	sens
C1(=C(NC(=O)NC1=O)N)N=O	2.003	1.004	0.994	1.95	1.023	0.911	emit
CC(C)C1=C(C2=NC(=O)N=C2C=C1)C(=O)O	2.135	1.072	0.991	2.124	1.103	0.928	emit
C1=C(C(=CC2=NC(=O)N=C21)OC(=O)O)N	2.155	1.096	0.968	2.101	1.078	0.95	emit
C1N=C2C(=CC=C2S1)N	2.361	1.205	0.959	1.964	0.996	0.972	emit
C1=CC=C2C(=C1)NC3=CC=CC(=O)C3=[N+]2[O-]	2.007	1.039	0.934	1.96	1.024	0.918	emit
C1=C[N+](=O)C(=O)C(=O)N1	1.963	1.027	0.916	2.027	1.039	0.953	emit

TABLE A.2: AL-ML candidates using loose NIR bounds

SMILES	S1	T1	FOM	S1 TDDFT	T1 TDDFT	FOM TDDFT	type
[CH2-]N1C[C@H]2CCN[C@H]2C1	1.236	1.223	0.989	1.09	1.083	0.994	sens
[CH2-]N1CCCC2C1CCCC2	1.121	1.101	0.983	1.156	1.114	0.964	sens
CCOC[C@@H]1CCN1[CH2-]	1.272	1.249	0.983	1.097	1.091	0.995	sens
CCC1=NC=C2N1CCN(C2)[CH2-]	1.174	1.142	0.973	1.161	1.156	0.996	sens
[B](C)C1=C2C(=CC=C1)N=CC=N2	1.317	1.276	0.969	1.177	1.177	1.0	sens
[CH2-]N(CCC#N)CCC1=CC=CC=C1	1.369	1.319	0.965	1.165	1.162	0.998	sens
CC(C1(CCN(CC1)[CH2-])O)(F)F	1.286	1.226	0.954	1.199	1.169	0.975	sens
C1(=C(NC(=O)NC1=O)N)N=O	2.003	1.004	0.994	1.95	1.023	0.911	emit
CC(C)C1=C(C2=NC(=O)N=C2C=C1)C(=O)O	2.135	1.072	0.991	2.124	1.103	0.928	emit
C1=C(C(=CC2=NC(=O)N=C21)OC(=O)O)N	2.155	1.096	0.968	2.101	1.078	0.95	emit
C1N=C2C(=CC=C2S1)N	2.361	1.205	0.959	1.964	0.996	0.972	emit
C1=CC=C2C(=C1)NC3=CC=CC(=O)C3=[N+]2[O-]	2.007	1.039	0.934	1.96	1.024	0.918	emit
CC1=NN2C(=NN(C2=C1N=O)C)C	1.685	0.879	0.921	1.918	1.002	0.918	emit
C1=C[N+](=O)C(=O)C(=O)N1	1.963	1.027	0.916	2.027	1.039	0.953	emit

A.3 ML model architectures

For DeepChem's GCN, a channel width of 64x64 was used for the graph convolutional layers, a channel width of 128 was used for the atom level dense layer, 75 atom features were created, a batch size of 100 was used, and a dropout of 0.2 was used.[147]

For DeepChem's MPNN, 75 features per atom were used, 14 features per atom pair were used, number of convolution depths in the corresponding hidden layer was 100, and a dropout of 0.2 was used.[148]

For Chemprop's MPNN, the hyperparameters used were: hidden size of 300, depth of 3, number of feed-forward layers of 2, and dropout of 0.[123]

A.4 MOPSSAM S1 comparison

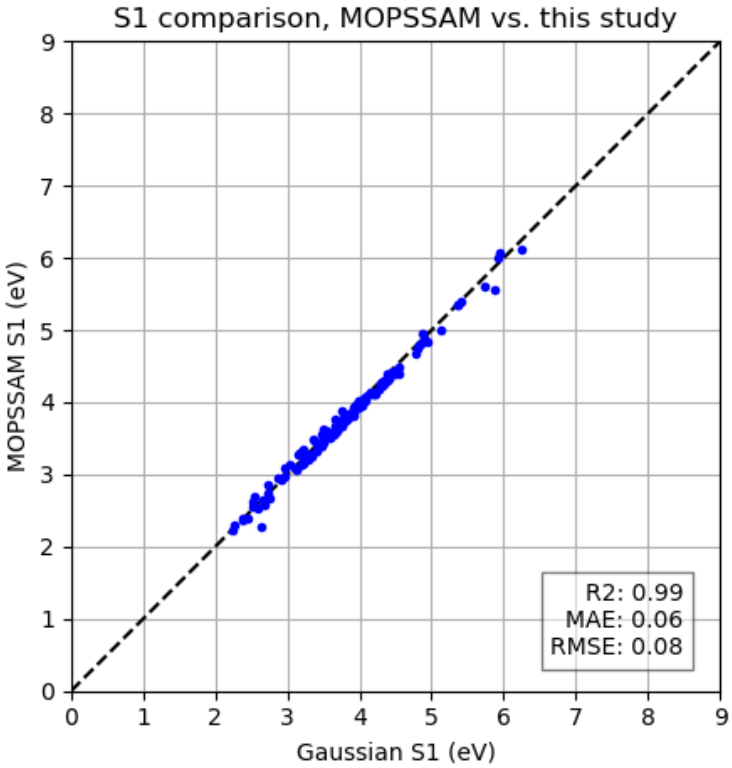


FIGURE A.2: Comparison of S1 energies calculated independently in this work vs. S1 energies calculated by Wilbraham et al. [**mopssam**], showing great agreement in results.

A.5 xTB-ML expanded training sets results

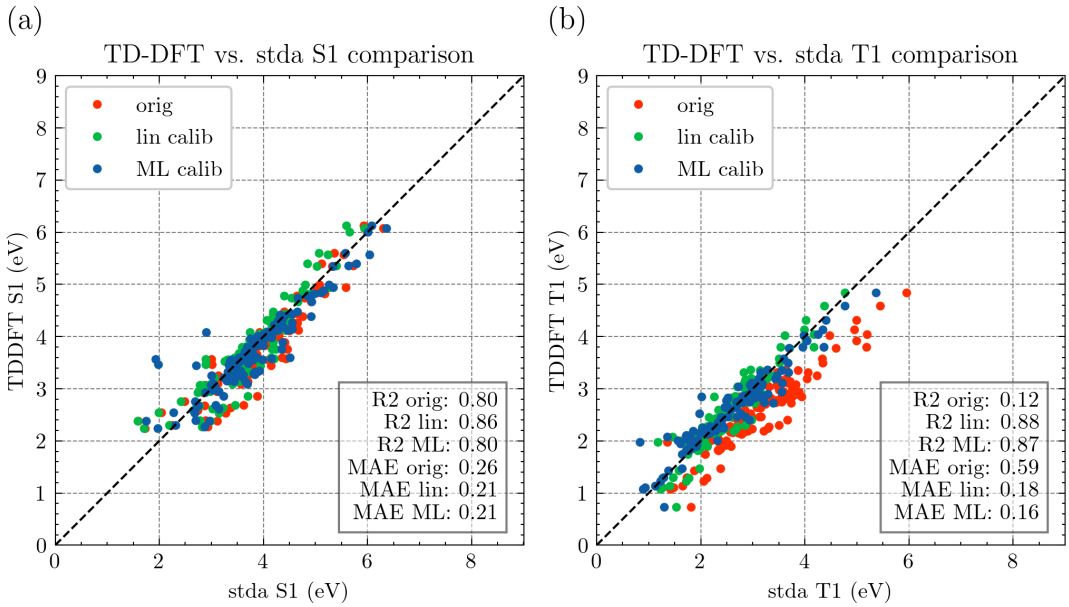


FIGURE A.3: Plot of xTB calibration of the 143 MOPSSAM molecules for (a) S1 and (b) T1 energies. Red dots are original data with no calibration, green dots are linearly calibrated data, and blue dots are calibrated with ML. Training data was the 22.5k molecules in SCOP-PCQC + SCOP-PCQC-lowS1 + QM-symex-10k, and test data was the 143 molecules shown here. Inlaid boxes show quantitative measurements of accuracy for original, linearly calibrated, and ML calibrated data.

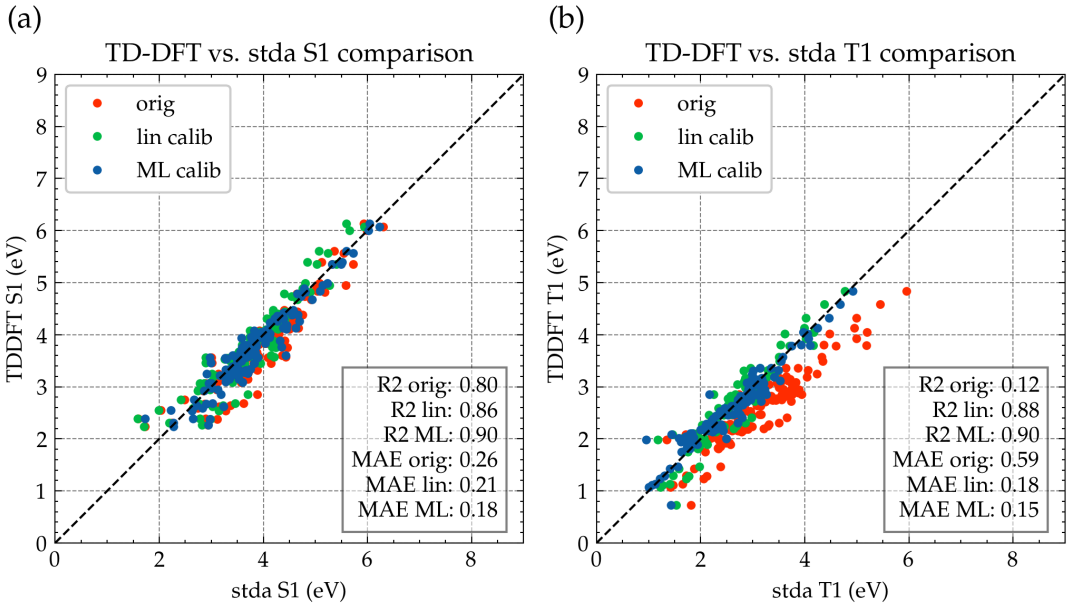


FIGURE A.4: Plot of xTB calibration of the 143 MOPSSAM molecules for (a) S1 and (b) T1 energies. Training data was the 78k molecules in SCOP-PCQC + SCOP-PCQC-ALS1 + QM-symex-10k, and test data was the 143 molecules shown here.

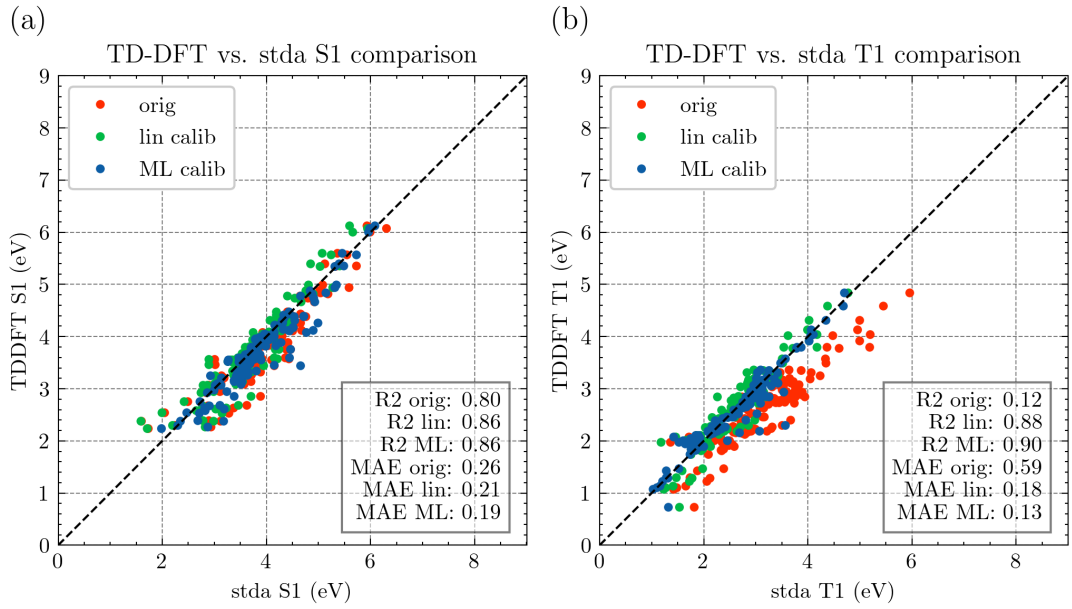


FIGURE A.5: Plot of xTB calibration of the 143 MOPSSAM molecules for (a) S1 and (b) T1 energies. Training data was the 127k molecules in SCOP-PCQC + SCOP-PCQC-ALT1 + QM-symex-10k, and test data was the 143 molecules shown here.

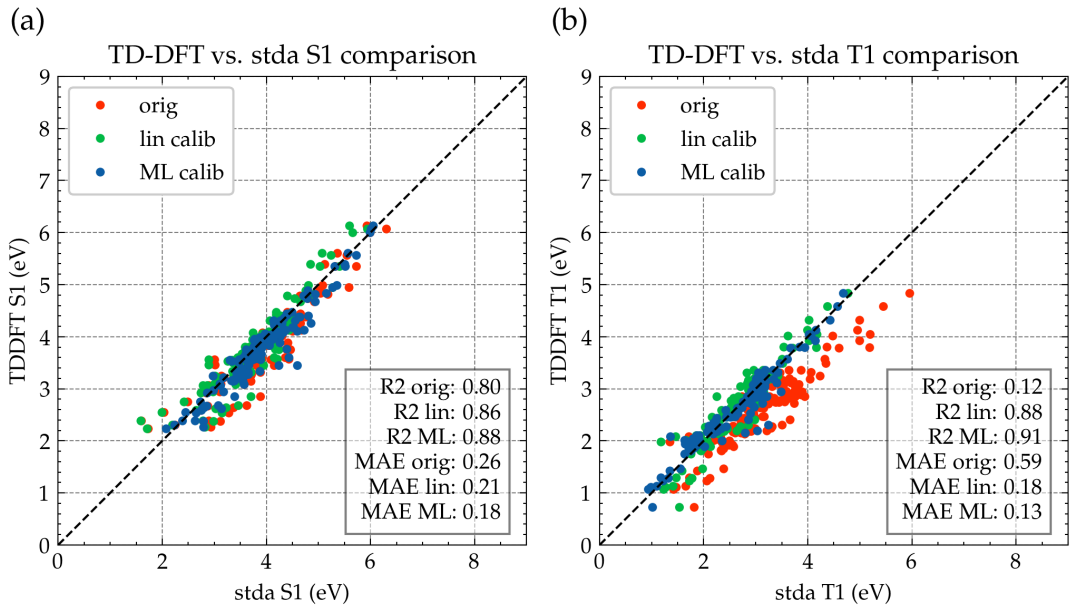


FIGURE A.6: Plot of xTB calibration of the 143 MOPSSAM molecules for (a) S1 and (b) T1 energies. Training data was the 182k molecules in SCOP-PCQC + SCOP-PCQC-ALS1 + SCOP-PCQC-ALT1 + QM-symex-10k, and test data was the 143 molecules shown here.

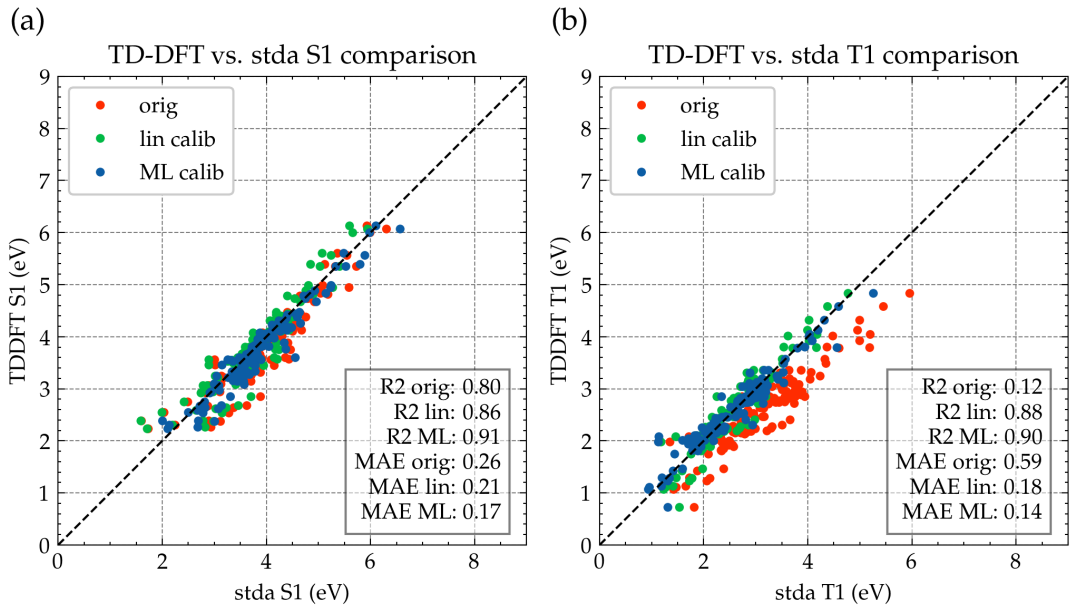


FIGURE A.7: Plot of xTB calibration of the 143 MOPSSAM molecules for (a) S1 and (b) T1 energies. Training data was the 138k molecules in SCOP-PCQC + QM-symex-10k + QM-symex, and test data was the 143 molecules shown here.

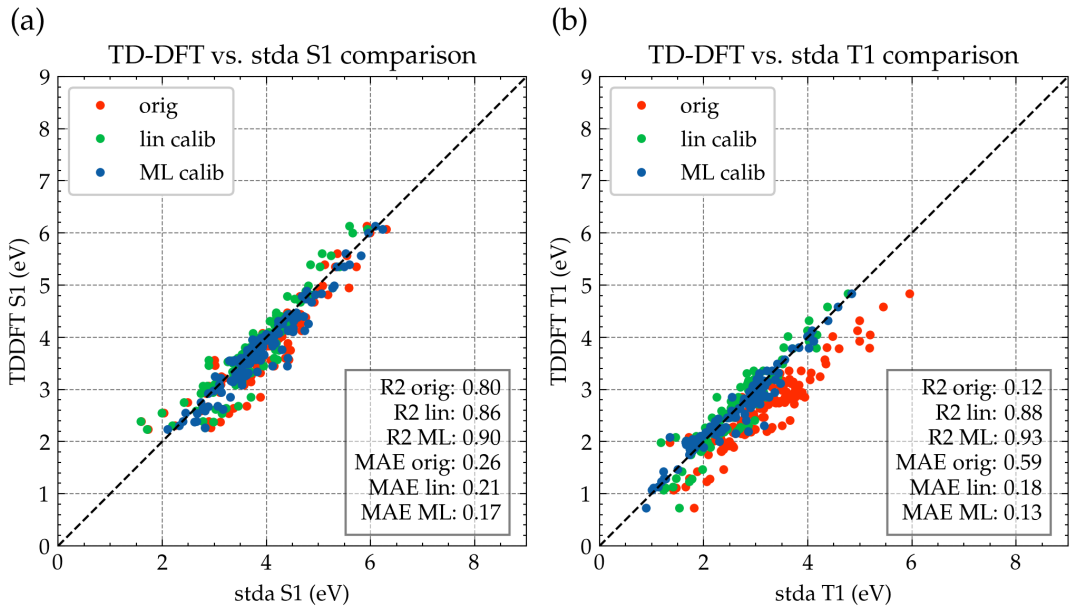


FIGURE A.8: Plot of xTB calibration of the 143 MOPSSAM molecules for (a) S1 and (b) T1 energies. Training data was the 301k molecules in SCOP-PCQC + SCOP-PCQC-ALS1 + SCOP-PCQC-ALT1 + QM-symex-10k + QM-symex, and test data was the 143 molecules shown here.

A.6 Applying xTB-ML to other functionals and methods

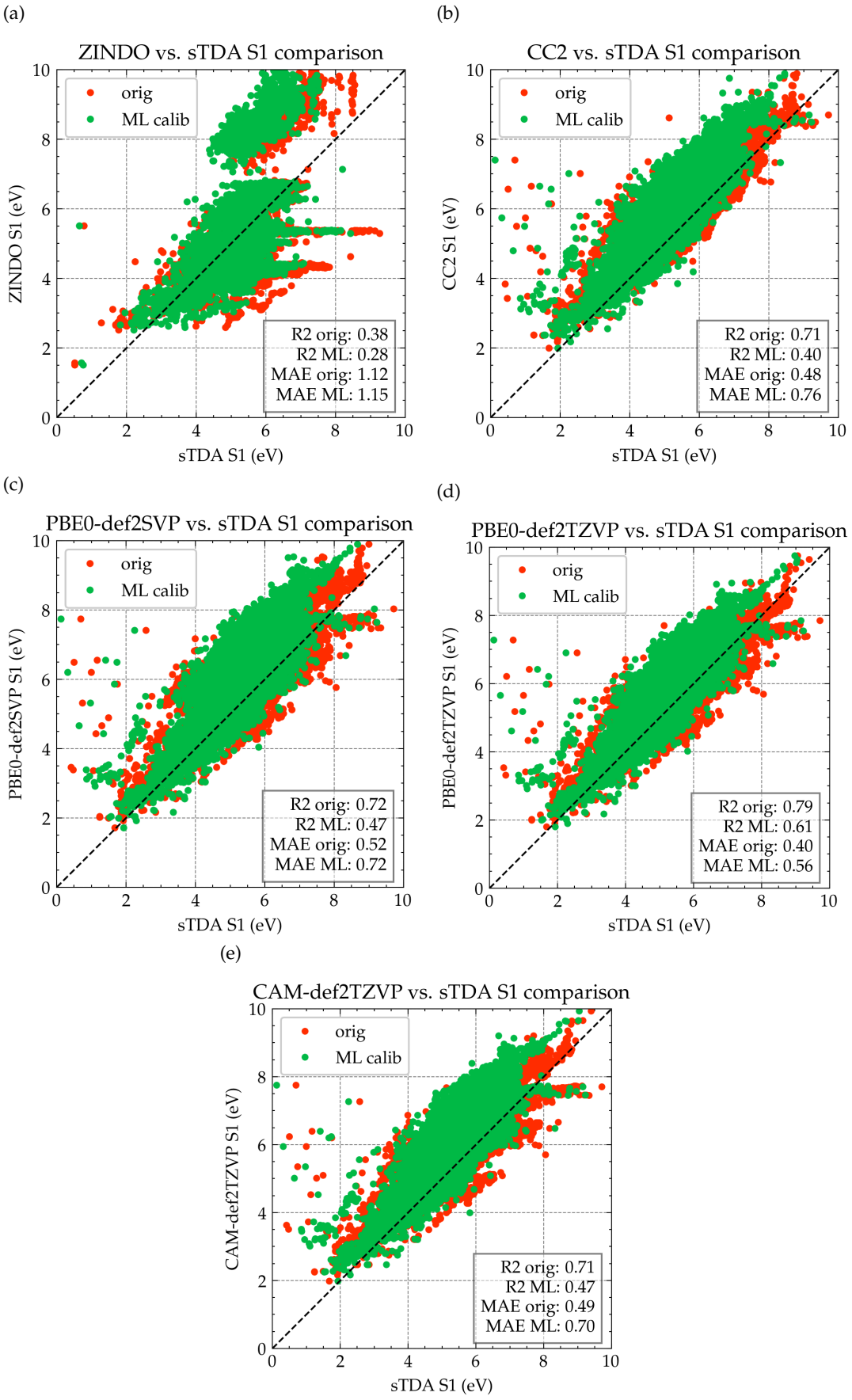


FIGURE A.9: Results of applying xTB-ML to other computational chemistry techniques: (a) ZINDO in QM7b and (b) CC2 in QM8, and TD-DFT functionals: (c) PBE0/def2-SVP in QM8, (d) PBE0/def2-TZVP in QM8, and (e) CAM-B3LYP/def2-TZVP in QM8.

Bibliography

(1) Imperial College Research Computing Service, DOI: 10.14469/hpc/2232.

(2) Kannan, N.; Vakeesan, D. *Renewable and Sustainable Energy Reviews* **2016**, *62*, Publisher: Elsevier Ltd, 1092–1105.

(3) Kabir, E.; Kumar, P.; Kumar, S.; Adelodun, A. A.; Kim, K.-H. *Renewable and Sustainable Energy Reviews* **2018**, *82*, 894–900.

(4) Solangi, K. H.; Islam, M. R.; Saidur, R.; Rahim, N. A.; Fayaz, H. *Renewable and Sustainable Energy Reviews* **2011**, *15*, Publisher: Pergamon, 2149–2163.

(5) Mekhilef, S.; Saidur, R.; Safari, A. *Renewable and Sustainable Energy Reviews* **2011**, *15*, Publisher: Pergamon, 1777–1790.

(6) Kim, H.; Park, E.; Kwon, S. J.; Ohm, J. Y.; Chang, H. J. *Renewable Energy* **2014**, *66*, 523–531.

(7) Burnett, D.; Barbour, E.; Harrison, G. P. *Renewable Energy* **2014**, *71*, 333–343.

(8) Fthenakis, V.; Mason, J. E.; Zweibel, K. *Energy Policy* **2009**, *37*, 387–399.

(9) Sharma, N. K.; Tiwari, P. K.; Sood, Y. R. *Renewable and Sustainable Energy Reviews* **2012**, *16*, 933–941.

(10) Liu, L. q.; Wang, Z. x.; Zhang, H. q.; Xue, Y. c. *Renewable and Sustainable Energy Reviews* **2010**, *14*, Publisher: Pergamon, 301–311.

(11) Bahadori, A.; Nwaoha, C. *Renewable and Sustainable Energy Reviews* **2013**, *18*, Publisher: Pergamon, 1–5.

(12) Dambhare, M. V.; Butey, B.; Moharil, S. V. *Journal of Physics: Conference Series* **2021**, *1913*, Publisher: IOP Publishing, 012053.

(13) Parida, B.; Iniyan, S.; Goic, R. *Renewable and Sustainable Energy Reviews* **2011**, *15*, 1625–1636.

(14) Shubbak, M. H. *Renewable and Sustainable Energy Reviews* **2019**, *115*, 109383.

(15) Gul, M.; Kotak, Y.; Muneer, T. *Energy Exploration & Exploitation* **2016**, *34*, Publisher: SAGE Publications Ltd STM, 485–526.

(16) Nayak, P. K.; Mahesh, S.; Snaith, H. J.; Cahen, D. *Nature Reviews Materials* **2019**, *4*, Bandiera_abtest: a Cg_type: Nature Research Journals Number: 4 Primary_atype: Research Publisher: Nature Publishing Group Subject_term: Semiconductors;Solar cells;Solar energy and photovoltaic technology Subject_term_id: semiconductors;solar-cells;solar-energy-and-photovoltaic-technology, 269–285.

(17) NREL Best Research-Cell Efficiency Chart, en, 2021.

(18) Jena, A. K.; Kulkarni, A.; Miyasaka, T. *Chemical Reviews* **2019**, *119*, Publisher: American Chemical Society, 3036–3103.

(19) Kim, J. Y.; Lee, J.-W.; Jung, H. S.; Shin, H.; Park, N.-G. *Chemical Reviews* **2020**, *120*, Publisher: American Chemical Society, 7867–7918.

(20) Lee, T. D.; Ebong, A. U. *Renewable and Sustainable Energy Reviews* **2017**, *70*, Publisher: Elsevier Ltd, 1286–1297.

(21) Kaur, N.; Singh, M.; Pathak, D.; Wagner, T.; Nunzi, J. M. *Synthetic Metals* **2014**, *190*, 20–26.

(22) Rühle, S. *Solar Energy* **2016**, *130*, 139–147.

(23) Day, J.; Senthilarasu, S.; Mallick, T. K. *Renewable Energy* **2019**, *132*, 186–205.

(24) Dimroth, F.; Kurtz, S. *MRS Bulletin* **2007**, *32*, 230–235.

(25) Philipps, S. P.; Bett, A. W.; Horowitz, K.; Kurtz, S. *Current Status of Concentrator Photovoltaic (CPV) Technology*; tech. rep. NREL/TP–5J00-65130, 1351597; 2015, NREL/TP–5J00–65130, 1351597.

(26) Ferry, D. K.; Goodnick, S. M.; Whiteside, V. R.; Sellers, I. R. *Journal of Applied Physics* **2020**, *128*, Publisher: American Institute of Physics, 220903.

(27) Peters, I. M.; Sofia, S.; Mailoa, J.; Buonassisi, T. *RSC Advances* **2016**, *6*, Publisher: The Royal Society of Chemistry, 66911–66923.

(28) McKenna, B.; Evans, R. C. *Advanced Materials* **2017**, *29*, 1606491–1606491.

(29) Huang, X.; Han, S.; Huang, W.; Liu, X. *Chemical Society Reviews* **2013**, *42*, Publisher: Royal Society of Chemistry, 173–201.

(30) Joubert, M.-F. *Optical Materials* **1999**, *11*, 181–203.

(31) Auzel, F. *Chemical Reviews* **2004**, *104*, Publisher: American Chemical Society, 139–174.

(32) Haase, M.; Schäfer, H. *Angewandte Chemie International Edition* **2011**, *50*, _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/anie.201005159, 5808–5829.

(33) LaCount, M. D.; Weingarten, D.; Hu, N.; Shaheen, S. E.; van de Lagemaat, J.; Rumbles, G.; Walba, D. M.; Lusk, M. T. *The Journal of Physical Chemistry A* **2015**, *119*, Publisher: American Chemical Society, 4009–4016.

(34) Wang, J.; Ming, T.; Jin, Z.; Wang, J.; Sun, L.-D.; Yan, C.-H. *Nature Communications* **2014**, *5*, Bandiera_abtest: a Cg_type: Nature Research Journals Number: 1 Primary_atype: Research Publisher: Nature Publishing Group Subject_term: Inorganic chemistry;Solar cells;Solar energy and photovoltaic technology Subject_term_id: inorganic-chemistry;solar-cells;solar-energy-and-photovoltaic-technology, 5669.

(35) Zhao, J.; Ji, S.; Guo, H. *RSC Advances* **2011**, *1*, Publisher: The Royal Society of Chemistry, 937–950.

(36) Simon, Y. C.; Weder, C. *Journal of Materials Chemistry* **2012**, *22*, Publisher: The Royal Society of Chemistry, 20817–20830.

(37) Couteau, C. *Contemporary Physics* **2018**, *59*, Publisher: Taylor & Francis _eprint: https://doi.org/10.1080/00107514.2018.1488463, 291–304.

(38) Ferro, S. M.; Wobben, M.; Ehrler, B. *Materials Horizons* **2021**, *8*, Publisher: The Royal Society of Chemistry, 1072–1083.

(39) Nozik, A. J. *Chemical Physics Letters* **2008**, *457*, 3–11.

(40) Smith, M. B.; Michl, J. *Chemical Reviews* **2010**, *110*, Publisher: American Chemical Society, 6891–6936.

(41) Sasikumar, D.; John, A. T.; Sunny, J.; Hariharan, M. *Chemical Society Reviews* **2020**, *49*, Publisher: The Royal Society of Chemistry, 6122–6140.

(42) Felter, K. M.; Grozema, F. C. *The Journal of Physical Chemistry Letters* **2019**, *10*, Publisher: American Chemical Society, 7208–7214.

(43) Ito, S.; Nagami, T.; Nakano, M. *Journal of Photochemistry and Photobiology C: Photochemistry Reviews* **2018**, *34*, 85–120.

(44) Feng, X.; Casanova, D.; Krylov, A. I. *The Journal of Physical Chemistry C* **2016**, *120*, Publisher: American Chemical Society, 19070–19077.

(45) Ni, W.; Gurzadyan, G. G.; Zhao, J.; Che, Y.; Li, X.; Sun, L. *The Journal of Physical Chemistry Letters* **2019**, *10*, Publisher: American Chemical Society, 2428–2433.

(46) Zirzlmeier, J.; Lehnherr, D.; Coto, P. B.; Chernick, E. T.; Casillas, R.; Basel, B. S.; Thoss, M.; Tykwinski, R. R.; Guldi, D. M. *Proceedings of the National Academy of Sciences* **2015**, *112*, Publisher: National Academy of Sciences Section: Physical Sciences, 5325–5330.

(47) Catti, L.; Narita, H.; Tanaka, Y.; Sakai, H.; Hasobe, T.; Tkachenko, N. V.; Yoshizawa, M. *Journal of the American Chemical Society* **2021**, *143*, Publisher: American Chemical Society, 9361–9367.

(48) Zirzlmeier, J.; Casillas, R.; Reddy, S. R.; Coto, P. B.; Lehnherr, D.; Chernick, E. T.; Papadopoulos, I.; Thoss, M.; Tykwinski, R. R.; Guldi, D. M. *Nanoscale* **2016**, *8*, Publisher: The Royal Society of Chemistry, 10113–10123.

(49) Walker, B. J.; Musser, A. J.; Beljonne, D.; Friend, R. H. *Nature Chemistry* **2013**, *5*, Bandiera_abtest: a Cg_type: Nature Research Journals Number: 12 Primary_atype: Research Publisher: Nature Publishing Group Subject_term: Chemical physics;Reaction kinetics and dynamics;Materials chemistry;Optical spectroscopy Subject_term_id: chemical-physics;kinetics-and-dynamics;materials-chemistry;spectroscopy, 1019–1024.

(50) Liu, H.; Wang, Z.; Wang, X.; Shen, L.; Zhang, C.; Xiao, M.; Li, X. *Journal of Materials Chemistry C* **2018**, *6*, Publisher: The Royal Society of Chemistry, 3245–3253.

(51) Stern, H. L.; Musser, A. J.; Gelinas, S.; Parkinson, P.; Herz, L. M.; Bruzek, M. J.; Anthony, J.; Friend, R. H.; Walker, B. J. *Proceedings of the National Academy of Sciences* **2015**, *112*, Publisher: National Academy of Sciences Section: Physical Sciences, 7656–7661.

(52) Fallon, K. J. et al. *Journal of the American Chemical Society* **2019**, *141*, Publisher: American Chemical Society, 13867–13876.

(53) El Bakouri, O.; Smith, J. R.; Ottosson, H. *Journal of the American Chemical Society* **2020**, *142*, Publisher: American Chemical Society, 5602–5617.

(54) Ye, C.; Zhou, L.; Wang, X.; Liang, Z. *Physical Chemistry Chemical Physics* **2016**, *18*, Publisher: The Royal Society of Chemistry, 10818–10835.

(55) Singh-Rachford, T. N.; Castellano, F. N. *Coordination Chemistry Reviews* **2010**, *254*, 2560–2573.

(56) Manna, M. K.; Shokri, S.; Wiederrecht, G. P.; Gosztola, D. J.; Ayitou, A. J.-L. *Chemical Communications* **2018**, *54*, Publisher: The Royal Society of Chemistry, 5809–5818.

(57) Bharmoria, P.; Bildirir, H.; Moth-Poulsen, K. *Chemical Society Reviews* **2020**, *49*, Publisher: The Royal Society of Chemistry, 6529–6554.

(58) Baluschev, S.; Yakutkin, V.; Miteva, T.; Avlasevich, Y.; Chernov, S.; Aleshchenkov, S.; Nelles, G.; Cheprakov, A.; Yasuda, A.; Müllen, K.; Wegner, G. *Angewandte Chemie International Edition* **2007**, *46*, _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/ani 7693–7696.

(59) Baluschev, S.; Yakutkin, V.; Miteva, T.; Wegner, G.; Roberts, T.; Nelles, G.; Yasuda, A.; Chernov, S.; Aleshchenkov, S.; Cheprakov, A. *New Journal of Physics* **2008**, *10*, Publisher: IOP Publishing, 013007.

(60) Deng, F.; Sommer, J. R.; Myahkostupov, M.; Schanze, K. S.; Castellano, F. N. *Chemical Communications* **2013**, *49*, Publisher: The Royal Society of Chemistry, 7406–7408.

(61) Mahboub, M.; Huang, Z.; Tang, M. L. *Nano Letters* **2016**, *16*, Publisher: American Chemical Society, 7169–7175.

(62) Haruki, R.; Sasaki, Y.; Masutani, K.; Yanai, N.; Kimizuka, N. *Chemical Communications* **2020**, *56*, Publisher: The Royal Society of Chemistry, 7017–7020.

(63) Radiunas, E.; Raišys, S.; Juršėnas, S.; Jozeliūnaitė, A.; Javorskis, T.; Šinkevičiūtė, U.; Orentas, E.; Kazlauskas, K. *Journal of Materials Chemistry C* **2020**, *8*, Publisher: The Royal Society of Chemistry, 5525–5534.

(64) Nienhaus, L.; Wu, M.; Geva, N.; Shepherd, J. J.; Wilson, M. W. B.; Bulović, V.; Van Voorhis, T.; Baldo, M. A.; Bawendi, M. G. *ACS Nano* **2017**, *11*, Publisher: American Chemical Society, 7848–7857.

(65) Nakata, M.; Shimazaki, T. *Journal of Chemical Information and Modeling* **2017**, *57*, Publisher: American Chemical Society, 1300–1308.

(66) Fang, C.; Oruganti, B.; Durbeej, B. *The Journal of Physical Chemistry A* **2014**, *118*, Publisher: American Chemical Society, 4157–4171.

(67) Westermayr, J.; Marquetand, P. *Machine Learning: Science and Technology* **2020**, *1*, Publisher: IOP Publishing, 043001.

(68) Stephens, P. J.; Devlin, F. J.; Chabalowski, C. F.; Frisch, M. J. *The Journal of Physical Chemistry* **1994**, *98*, Publisher: American Chemical Society, 11623–11627.

(69) Zhang, I. Y.; Wu, J.; Xu, X. *Chemical Communications* **2010**, *46*, Publisher: The Royal Society of Chemistry, 3057–3070.

(70) Becke, A. D. *Physical Review A* **1988**, *38*, Publisher: American Physical Society, 3098–3100.

(71) Lee, C.; Yang, W.; Parr, R. G. *Physical Review B* **1988**, *37*, Publisher: American Physical Society, 785–789.

(72) Frisch, M. J.; Pople, J. A.; Binkley, J. S. *The Journal of Chemical Physics* **1984**, *80*, Publisher: American Institute of Physics, 3265–3269.

(73) Hehre, W. J.; Ditchfield, R.; Pople, J. A. *The Journal of Chemical Physics* **1972**, *56*, Publisher: American Institute of Physics, 2257–2261.

(74) Kendall, R. A.; Dunning, T. H.; Harrison, R. J. *The Journal of Chemical Physics* **1992**, *96*, Publisher: American Institute of Physics, 6796–6806.

(75) Weigend, F.; Ahlrichs, R. *Physical Chemistry Chemical Physics* **2005**, *7*, Publisher: The Royal Society of Chemistry, 3297–3305.

(76) Jacquemin, D.; Mennucci, B.; Adamo, C. *Physical Chemistry Chemical Physics* **2011**, *13*, Publisher: The Royal Society of Chemistry, 16987–16998.

(77) Levine, B. G.; Martínez, T. J. *Annual Review of Physical Chemistry* **2007**, *58*, _eprint: https://doi.org/10.1146/annurev.physchem.57.032905.104612, 613–634.

(78) Hirata, S.; Head-Gordon, M. *Chemical Physics Letters* **1999**, *314*, 291–299.

(79) Ruddigkeit, L.; van Deursen, R.; Blum, L. C.; Reymond, J.-L. *Journal of Chemical Information and Modeling* **2012**, *52*, Publisher: American Chemical Society, 2864–2875.

(80) Kim, S.; Thiessen, P. A.; Bolton, E. E.; Chen, J.; Fu, G.; Gindulyte, A.; Han, L.; He, J.; He, S.; Shoemaker, B. A.; Wang, J.; Yu, B.; Zhang, J.; Bryant, S. H. *Nucleic Acids Research* **2016**, *44*, D1202–D1213.

(81) Smith, D. G. A.; Altarawy, D.; Burns, L. A.; Welborn, M.; Naden, L. N.; Ward, L.; Ellis, S.; Pritchard, B. P.; Crawford, T. D. *WIREs Computational Molecular Science* **2021**, *11*, _eprint: https://wires.onlinelibrary.wiley.com/doi/pdf/10.1002/wcms.1491, e1491.

(82) Smith, J. S.; Isayev, O.; Roitberg, A. E. *Chemical Science* **2017**, *8*, Publisher: The Royal Society of Chemistry, 3192–3203.

(83) Ramakrishnan, R.; Dral, P. O.; Rupp, M.; von Lilienfeld, O. A. *Scientific Data* **2014**, *1*, Number: 1 Publisher: Nature Publishing Group, 140022.

(84) Montavon, G.; Rupp, M.; Gobre, V.; Vazquez-Mayagoitia, A.; Hansen, K.; Tkatchenko, A.; Müller, K.-R.; Lilienfeld, O. A. v. *New Journal of Physics* **2013**, *15*, Publisher: IOP Publishing, 095003.

(85) Rupp, M.; Tkatchenko, A.; Müller, K.-R.; von Lilienfeld, O. A. *Physical Review Letters* **2012**, *108*, Publisher: American Physical Society, 058301.

(86) Ridley, J.; Zerner, M. *Theoretica chimica acta* **1973**, *32*, 111–134.

(87) Tkatchenko, A.; DiStasio, R. A.; Car, R.; Scheffler, M. *Physical Review Letters* **2012**, *108*, Publisher: American Physical Society, 236402.

(88) Perdew, J. P.; Ernzerhof, M.; Burke, K. *The Journal of Chemical Physics* **1996**, *105*, Publisher: American Institute of Physics, 9982–9985.

(89) Hedin, L. *Physical Review* **1965**, *139*, Publisher: American Physical Society, A796–A823.

(90) Blum, L. C.; Reymond, J.-L. *Journal of the American Chemical Society* **2009**, *131*, Publisher: American Chemical Society, 8732–8733.

(91) Ramakrishnan, R.; Hartmann, M.; Tapavicza, E.; von Lilienfeld, O. A. *The Journal of Chemical Physics* **2015**, *143*, Publisher: American Institute of Physics, 084111.

(92) Hättig, C.; Weigend, F. *The Journal of Chemical Physics* **2000**, *113*, Publisher: American Institute of Physics, 5154–5161.

(93) Abreha, B. G.; Agarwal, S.; Foster, I.; Blaiszik, B.; Lopez, S. A. *The Journal of Physical Chemistry Letters* **2019**, *10*, Publisher: American Chemical Society, 6835–6841.

(94) Zhao, Y.; Truhlar, D. G. *Theoretical Chemistry Accounts* **2008**, *120*, 215–241.

(95) Liang, J.; Ye, S.; Dai, T.; Zha, Z.; Gao, Y.; Zhu, X. *Scientific Data* **2020**, *7*, Number: 1 Publisher: Nature Publishing Group, 400.

(96) Liang, J.; Xu, Y.; Liu, R.; Zhu, X. *Scientific Data* **2019**, *6*, Number: 1 Publisher: Nature Publishing Group, 213.

(97) Schwilk, M.; Tahchieva, D. N.; von Lilienfeld, O. A. *arXiv:2004.10600 [physics]* **2020**, arXiv: 2004.10600.

(98) Véril, M.; Scemama, A.; Caffarel, M.; Lipparini, F.; Boggio-Pasqua, M.; Jacquemin, D.; Loos, P.-F. *WIREs Computational Molecular Science* **2021**, *11*, _eprint: https://wires.onlinelibrary.wi e1517.

(99) Spiegelman, F.; Tarrat, N.; Cuny, J.; Dontot, L.; Posenitskiy, E.; Martí, C.; Simon, A.; Rapacioli, M. *Advances in Physics: X* **2020**, *5*, Publisher: Taylor & Francis _eprint: https://doi.org/10.1080/23746149.2019.1710252, 1710252.

(100) Elstner, M.; Porezag, D.; Jungnickel, G.; Elsner, J.; Haugk, M.; Frauenheim, T.; Suhai, S.; Seifert, G. *Physical Review B* **1998**, *58*, Publisher: American Physical Society, 7260–7268.

(101) Bannwarth, C.; Caldeweyher, E.; Ehlert, S.; Hansen, A.; Pracht, P.; Seibert, J.; Spicher, S.; Grimme, S. *WIREs Computational Molecular Science* **2021**, *11*, _eprint: https://wires.onlinelibrary.wiley.com/doi/pdf/10.1002/wcms.1493, e1493.

(102) Grimme, S.; Bannwarth, C.; Shushkov, P. *Journal of Chemical Theory and Computation* **2017**, *13*, Publisher: American Chemical Society, 1989–2009.

(103) Bannwarth, C.; Ehlert, S.; Grimme, S. *Journal of Chemical Theory and Computation* **2019**, *15*, Publisher: American Chemical Society, 1652–1671.

(104) Grimme, S. *The Journal of Chemical Physics* **2013**, *138*, Publisher: American Institute of Physics, 244104.

(105) Grimme, S.; Bannwarth, C. *The Journal of Chemical Physics* **2016**, *145*, Publisher: American Institute of Physics, 054103.

(106) Wiebeler, C.; Schapiro, I. *Molecules* **2019**, *24*, Number: 9 Publisher: Multidisciplinary Digital Publishing Institute, 1720.

(107) Wilbraham, L.; Berardo, E.; Turcani, L.; Jelfs, K. E.; Zwijnenburg, M. A. *Journal of Chemical Information and Modeling* **2018**, *58*, Publisher: American Chemical Society, 2450–2459.

(108) Batra, K.; Zahn, S.; Heine, T. *Advanced Theory and Simulations* **2020**, *3*, _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/adts.201900192, 1900192.

(109) Liu, Y.; Xu, J.; Han, L.; Liu, Q.; Yang, Y.; Li, Z.; Lu, Z.; Zhang, H.; Guo, T.; Liu, Q. *Interdisciplinary Sciences: Computational Life Sciences* **2021**, *13*, 140–146.

(110) Yanai, T.; Tew, D. P.; Handy, N. C. *Chemical Physics Letters* **2004**, *393*, 51–57.

(111) Wilbraham, L.; Sprick, R. S.; Jelfs, K. E.; Zwijnenburg, M. A. *Chemical Science* **2019**, *10*, Publisher: The Royal Society of Chemistry, 4973–4984.

(112) Wilbraham, L.; Smajli, D.; Heath-Apostolopoulos, I.; Zwijnenburg, M. A. *Communications Chemistry* **2020**, *3*, Number: 1 Publisher: Nature Publishing Group, 1–9.

(113) Heath-Apostolopoulos, I.; Wilbraham, L.; Zwijnenburg, M. A. *Faraday Discussions* **2019**, *215*, Publisher: The Royal Society of Chemistry, 98–110.

(114) Heath-Apostolopoulos, I.; Vargas-Ortiz, D.; Wilbraham, L.; Jelfs, K. E.; Zwijnenburg, M. A. *Sustainable Energy & Fuels* **2021**, *5*, Publisher: The Royal Society of Chemistry, 704–719.

(115) Musil, F.; Grisafi, A.; Bartók, A. P.; Ortner, C.; Csányi, G.; Ceriotti, M. *Chemical Reviews* **2021**, Publisher: American Chemical Society, DOI: 10.1021/acs.chemrev.1c00021.

(116) Ward, L. et al. *Computational Materials Science* **2018**, *152*, 60–69.

(117) Dral, P. O. *Journal of Computational Chemistry* **2019**, *40*, _eprint: https://onlinelibrary.wiley.com/doi/ 2339–2347.

(118) Wu, Z.; Ramsundar, B.; N. Feinberg, E.; Gomes, J.; Geniesse, C.; S. Pappu, A.; Leswing, K.; Pande, V. *Chemical Science* **2018**, *9*, Publisher: Royal Society of Chemistry, 513–530.

(119) Schütt, K. T.; Arbabzadah, F.; Chmiela, S.; Müller, K. R.; Tkatchenko, A. *Nature Communications* **2017**, *8*, Bandiera_abtest: a Cc_license_type: cc_by Cg_type: Nature Research Journals Number: 1 Primary_atype: Research Publisher: Nature Publishing Group Subject_term: Applied mathematics;Computational chemistry;Physical chemistry;Scientific data Subject_term_id: applied-mathematics;computational-chemistry;physical-chemistry;scientific-data, 13890.

(120) Westermayr, J.; Faber, F. A.; Christensen, A. S.; Lilienfeld, O. A. v.; Marquetand, P. *Machine Learning: Science and Technology* **2020**, *1*, Publisher: IOP Publishing, 025009.

(121) Jiang, D.; Wu, Z.; Hsieh, C.-Y.; Chen, G.; Liao, B.; Wang, Z.; Shen, C.; Cao, D.; Wu, J.; Hou, T. *Journal of Cheminformatics* **2021**, *13*, 12.

(122) Gilmer, J.; Schoenholz, S. S.; Riley, P. F.; Vinyals, O.; Dahl, G. E. In *International Conference on Machine Learning*; ISSN: 2640-3498, PMLR: 2017, pp 1263–1272.

(123) Yang, K.; Swanson, K.; Jin, W.; Coley, C.; Eiden, P.; Gao, H.; Guzman-Perez, A.; Hopper, T.; Kelley, B.; Mathea, M.; Palmer, A.; Settels, V.; Jaakkola, T.; Jensen, K.; Barzilay, R. *Journal of Chemical Information and Modeling* **2019**, *59*, Publisher: American Chemical Society, 3370–3388.

(124) O'Boyle, N. M.; Banck, M.; James, C. A.; Morley, C.; Vandermeersch, T.; Hutchison, G. R. *Journal of Cheminformatics* **2011**, *3*, 33.

(125) Perdew, J. P.; Burke, K.; Ernzerhof, M. *Physical Review Letters* **1996**, *77*, Publisher: American Physical Society, 3865–3868.

(126) Pronobis, W.; Schütt, K. T.; Tkatchenko, A.; Müller, K.-R. *The European Physical Journal B* **2018**, *91*, 178.

(127) Ghosh, K.; Stuke, A.; Todorović, M.; Jørgensen, P. B.; Schmidt, M. N.; Vehtari, A.; Rinke, P. *Advanced Science* **2019**, *6*, _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/ad 1801367.

(128) Kang, B.; Seok, C.; Lee, J. *Journal of Chemical Information and Modeling* **2020**, *60*, Publisher: American Chemical Society, 5984–5994.

(129) Ramakrishnan, R.; Dral, P. O.; Rupp, M.; von Lilienfeld, O. A. *Journal of Chemical Theory and Computation* **2015**, *11*, Publisher: American Chemical Society, 2087–2096.

(130) Pollice, R.; Friederich, P.; Lavigne, C.; Gomes, G. d. P.; Aspuru-Guzik, A. *Matter* **2021**, *4*, 1654–1682.

(131) Hu, L.; Wang, X.; Wong, L.; Chen, G. *The Journal of Chemical Physics* **2003**, *119*, Publisher: American Institute of Physics, 11501–11507.

(132) Sun, J.; Wu, J.; Song, T.; Hu, L.; Shan, K.; Chen, G. *The Journal of Physical Chemistry A* **2014**, *118*, Publisher: American Chemical Society, 9120–9131.

(133) Yang, G.; Wu, J.; Chen, S.; Zhou, W.; Sun, J.; Chen, G. *The Journal of Chemical Physics* **2018**, *148*, Publisher: American Institute of Physics, 241738.

(134) Wang, X.; Wong, L.; Hu, L.; Chan, C.; Su, Z.; Chen, G. *The Journal of Physical Chemistry A* **2004**, *108*, Publisher: American Chemical Society, 8514–8525.

(135) Li, H.; Shi, L.; Zhang, M.; Su, Z.; Wang, X.; Hu, L.; Chen, G. *The Journal of Chemical Physics* **2007**, *126*, Publisher: American Institute of Physics, 144101.

(136) Seung, H. S.; Opper, M.; Sompolinsky, H. In *Proceedings of the fifth annual workshop on Computational learning theory*, Association for Computing Machinery: New York, NY, USA, 1992, pp 287–294.

(137) Gubaev, K.; Podryabinkin, E. V.; Shapeev, A. V. *The Journal of Chemical Physics* **2018**, *148*, Publisher: American Institute of Physics, 241727.

(138) Kunkel, C.; Margraf, J. T.; Chen, K.; Oberhofer, H.; Reuter, K. *Nature Communications* **2021**, *12*, Bandiera_abtest: a Cc_license_type: cc_by Cg_type: Nature Research Journals Number: 1 Primary_atype: Research Publisher: Nature Publishing Group Subject_term: Computational chemistry;Organic molecules in materials science Subject_term_id: computational-chemistry;organic-molecules-in-materials-science, 2422.

(139) Smith, J. S.; Nebgen, B.; Lubbers, N.; Isayev, O.; Roitberg, A. E. *The Journal of Chemical Physics* **2018**, *148*, Publisher: American Institute of Physics, 241733.

(140) Gómez-Bombarelli, R. et al. *Nature Materials* **2016**, *15*, 1120–1127.

(141) McInnes, L.; Healy, J.; Melville, J. *arXiv:1802.03426 [cs, stat]* **2020**, arXiv: 1802.03426.

(142) shomikverma/AL_ML_data: Identified molecular chromophores for photon conversion processes using active machine learning for energy prediction. en.

(143) Chen, T.; Zheng, L.; Yuan, J.; An, Z.; Chen, R.; Tao, Y.; Li, H.; Xie, X.; Huang, W. *Scientific Reports* **2015**, *5*, Bandiera_abtest: a Cc_license_type: cc_by Cg_type: Nature Research Journals Number: 1 Primary_atype: Research Publisher: Nature Publishing Group Subject_term: Computational chemistry;Excited states;Photochemistry Subject_term_id: computational-chemistry;excited-states;photochemistry, 10923.

(144) Jensen, J. H. *Chemical Science* **2019**, *10*, Publisher: The Royal Society of Chemistry, 3567–3572.

(145) Henault, E. S.; Rasmussen, M. H.; Jensen, J. H. *PeerJ Physical Chemistry* **2020**, *2*, Publisher: PeerJ Inc., e11.

(146) Ramsundar, B.; Eastman, P.; Walters, P.; Pande, V., *Deep learning for the life sciences: applying deep learning to genomics, microscopy, drug discovery and more*, First edition, OCLC: on1051083869; O'Reilly Media: Sebastopol, CA, 2019.

(147) Duvenaud, D. K.; Maclaurin, D.; Iparraguirre, J.; Bombarell, R.; Hirzel, T.; Aspuru-Guzik, A.; Adams, R. P. In *Advances in Neural Information Processing Systems*, Curran Associates, Inc.: 2015; Vol. 28.

(148) Vinyals, O.; Bengio, S.; Kudlur, M. *arXiv:1511.06391 [cs, stat]* **2016**, arXiv: 1511.06391.

(149) Elsevier Developer Portal.

(150) Swain, M. C.; Cole, J. M. *Journal of Chemical Information and Modeling* **2016**, *56*, Publisher: American Chemical Society, 1894–1904.

(151) PUG REST.

(152) Hendrickson, J. B.; Huang, P.; Toczko, A. G. *Journal of Chemical Information and Computer Sciences* **1987**, *27*, Publisher: American Chemical Society, 63–67.

(153) Generate a single conformer — Open Babel 3.0.1 documentation.

(154) Yoshikawa, N.; Hutchison, G. R. *Journal of Cheminformatics* **2019**, *11*, 49.

(155) Campello, R. J. G. B.; Moulavi, D.; Sander, J. In *Advances in Knowledge Discovery and Data Mining*, ed. by Pei, J.; Tseng, V. S.; Cao, L.; Motoda, H.; Xu, G., Springer: Berlin, Heidelberg, 2013, pp 160–172.

(156) The hdbscan Clustering Library — hdbscan 0.8.1 documentation.

(157) Wilbraham, L. molZ, original-date: 2020-12-06T13:18:42Z, 2021.

(158) Pracht, P.; Bohle, F.; Grimme, S. *Physical Chemistry Chemical Physics* **2020**, *22*, Publisher: The Royal Society of Chemistry, 7169–7192.

(159) Li, S.; Zhou, Z.; Upadhayay, A.; Shaikh, O.; Freitas, S.; Park, H.; Wang, Z. J.; Routray, S.; Hull, M.; Chau, D. H. In *Proceedings of the 29th ACM International Conference on Information & Knowledge Management*, Association for Computing Machinery: New York, NY, USA, 2020, pp 3071–3076.

(160) Gao, X.; Bai, S.; Fazzi, D.; Niehaus, T.; Barbatti, M.; Thiel, W. *Journal of Chemical Theory and Computation* **2017**, *13*, Publisher: American Chemical Society, 515–524.