Virtual Excited State Reference for the Discovery of Electronic Materials Database (VERDE Materials DB): An open-access resource for ground and excited state properties of organic molecules

Biruk G. Abreha[1], Snigdha Agarwal[1], Ian Foster[2,3], Ben Blaiszik[2,3], Steven A. Lopez[1,*]

[1] Northeastern University, Boston, Massachusetts, United States
[2] Argonne National Laboratory, Lemont, IL, United States
[3] University of Chicago, Chicago, IL, United States

*s.lopez@northeastern.edu

ABSTRACT:

This article announces **VERDE materials DB**, the first database to include downloadable excited-state structures (S_0, S_1, T_1) and photophysical properties. **VERDE materials DB** is searchable, open-access via www.verdedb.org, and focused on light-responsive π-conjugated organic molecules with applications in green chemistry, organic solar cells, and organic redox flow batteries. It includes results of our active and past virtual screening studies; to date, more than 11,000 density functional theory (DFT) calculations have been performed on 1,282 molecules to obtain frontier molecular orbitals, and photophysical properties, including excitation energies, dipole moments, and redox potentials. To improve community access, we have made **VERDE materials DB** available via an integration with the Materials Data Facility. We are leveraging **VERDE materials DB** to train machine learning algorithms to identify new materials and structure-property relationships between molecular ground- and excited-states. We present a case-study involving photoaffinity labels, where we identify new diazirine-based photoaffinity labels with optimal photostabilities.

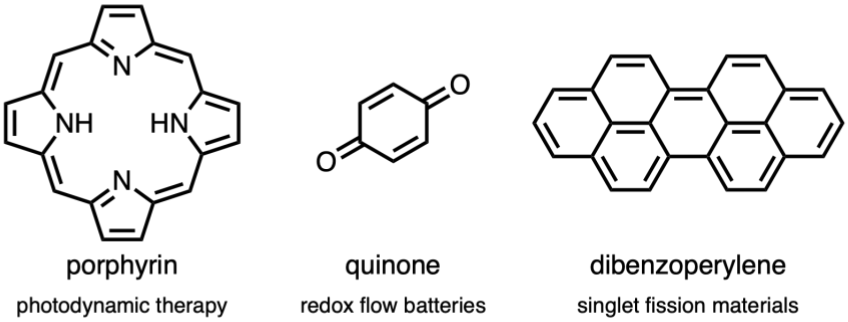
Introduction

Approximately 50,000 exajoules of harvestable solar energy reach the Earth each year, far exceeding the 400 exajoule total global energy consumption in 2016.[1, 2] The most recent inorganic photovoltaic devices are able to capture this energy with power conversion efficiencies (PCEs) exceeding 47%; in contrast, the highest confirmed PCEs for organic photovoltaics (OPVs) recently surpassed 16%.[3, 4] The solar day-night cycle interrupts solar energy conversion, thus making solar energy storage an equal priority for renewable energy research. Organic redox flow batteries, which use dissolved, electronically-active organic materials (*e.g.*, quinone and anthraquinone derivatives), have shown potential for large-scale energy storage.[5, 6] While inorganic materials have higher PCEs and battery efficiencies, their relatively high cost makes commercialization difficult and often requires subsidies.[7] In addition, organic energy materials provide green alternatives to commonly used inorganic materials.[5, 8, 9] Renewable solar energy has led to increasingly sustainable chemical reactions by eliminating the need for Earth-rare and organometallic catalysts for powerful organic transformations. The low-cost of organic materials combined with their straightforward processability and tunability suggests that sustainable next-generation devices and reactions, including singlet fission materials, organic photoredox catalyst-substrate pairs, and photoaffinity labels, are possible, but have yet to be discovered.[10-12]

Organic chromophores of broadest general interest absorb UV, visible, or near-IR light, depending on the application. They are typically π-conjugated and often feature aromatic moieties.

Porphyrins, quinones, and dibenzoperylenes are representative examples and are shown in Scheme 1. The vast number of possible chromophores compounded by the substitution patterns and possible functional groups makes the number of possible accessible organic molecules approach 10^{23}.[13]

Scheme 1. Examples of π-conjugated and aromatic molecule and their applications.



Experimental determination of molecular structure and properties is extremely expensive in terms of human time and chemical costs. An emerging approach involving quantum mechanical (QM) calculations combined with data-driven techniques (*e.g.*, machine learning) has facilitated the navigation of chemical space and 'smart' searches of chemical space. The QM calculations are typically density functional theory (DFT) and provide optimized geometries and electronic structures at reasonable cost. Machine learning (ML) algorithms—especially neural nets—require large datasets that are relatively rare in academia and proprietary in industry. The QM/ML approach allows scientists to determine the structures and properties of molecules and materials relatively quickly with high performance computing (HPC) resources and large datasets can be compiled. Indeed, the Harvard Clean Energy Project (CEP) contains an open-access dataset of 2.3 million candidate organic photovoltaic (OPV) materials and their predicted highest occupied molecular orbitals (HOMOs), lowest unoccupied molecular orbitals (LUMOs), and corresponding short-circuit current densities, open circuit voltages, and power conversion efficiencies, computed with the Scharber model.[14-16]

Databases of computed physicochemical properties of organic compounds can reveal trends in properties and help establish QSPR models which guide the rational design of new materials. Existing databases of organic compounds highlight their utility. GDB-13 enumerates 970 million synthetically-accessible, organic molecules containing up to 13 heavy atoms (C, N, O, S, Cl).[17] The QM7 dataset provides, for the subset of GDB-13 containing up to seven heavy atoms, Coulomb matrices and atomization energies for 7,165 organic molecules and was successfully used to train a nonlinear regression machine learning model to predict atomization energies based on molecular geometry and nuclear charges.[18] QM7b extends QM7 with 13 additional properties, such as HOMO and LUMO energies, polarizabilities, and excitation energies for 7,211 organic molecules.[19] Montavon *et al.* used this dataset to train multi-task deep neural network to predict, with reasonable accuracy, these additional properties using Coulomb matrices as descriptors. GDB-17, which extends GDB-13 to organic molecules containing up to 17 heavy atoms, enumerates 166 billion molecules. Von Lilienfeld *et al.* constructed the QM9 dataset, the subset of GDB-17 containing up to 9 heavy atoms, featuring ground state geometries, dipole moments, polarizabilities, enthalpies, and free energies for approximately 134,000 molecules.[20] QM9 has been used by many groups to construct neural networks to predict, with DFT-level accuracy, molecular properties at a relatively low computational cost.[21-23]

These databases of computed chemical properties have proven to be useful in the material discovery process.[21-24] Organic electronics function when constituent materials are in non-equilibrium states (*e.g.*, oxidized or photoexcited). The non-equilibrium structures are critically important to understanding the properties of these materials yet absent from current open-source large databases. However, current open-access databases, including QM7, QM9, and CEP, do not include excited state properties, such as structures and transition energies, which we have shown to be useful in understanding photophysical properties and photochemical reaction mechanisms. As shown in our case study below, the optimized excited-state structures of diazirines provide important clues about the photostabilities of diazirines. The open-access nature of **VERDE materials DB** means that research groups everywhere can discover new materials and infer fundamental structure-property relationships.

Results/Discussion

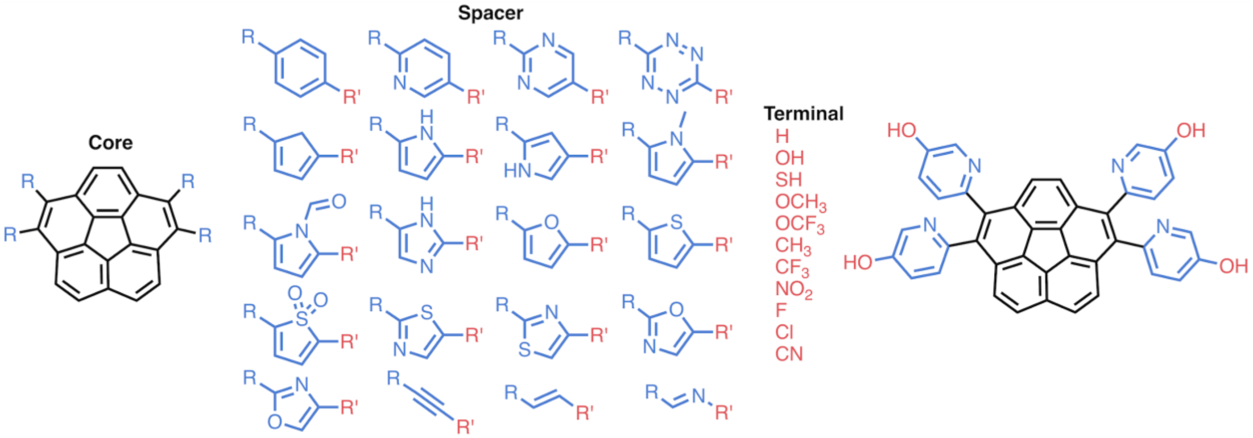
Organization of the data into the VERDE materials DB

This manuscript introduces the Virtual Excited State Reference for the Discovery of Electronic materials database (**VERDE materials DB**). We make VERDE materials DB and the associated data (i.e., calculation output files, and derived calculation results) openly available through an integration with data services provided by the Materials Data Facility (MDF).[25] We have implemented a flow where data supporting **VERDE materials DB** are published to MDF as they become available, important information about each calculation is automatically extracted and loaded into a search index, and the associated data are discoverable via advanced search capabilities, including partial matching and range queries. The open access nature of this database is meant to speed the discovery of new materials through simplified collection of data upon which machine learning and other analyses may be performed. We see opportunities in the future to leverage other data services, like the Data and Learning Hub for Science (DLHub), to act as a central repository of machine learning models derived from this database, to enable users to run models on new data, to benchmark and compare models, and to directly link these models to training data from **VERDE materials DB**.

As such, **VERDE materials DB** will meet a substantial need from the experimental and theoretical communities developing sustainable materials for OPVs, organic field-effect transistors, and green chemistry (*e.g.*, photoredox catalysts). **VERDE materials DB** is the first containing extensive ground and excited state, DFT-optimized geometries and thermochemical calculations for organic materials. The computed electronic states include the S_0, S_1, T_1, and radical cation states (see Methods for computational methodology). Further, **VERDE materials DB** includes properties computed from these DFT calculations such as redox potentials, 0-0 transition energies (interchangeable with E^{0-0} throughout the manuscript), and ionization potentials.[26, 27] E^{0-0} requires the optimization of the chromophore in a given excited state. Computations of vertical excitation energies are shown to be functional-dependent because the functionals can be overfit for classes of chromophores. Given the vast molecular diversity in **VERDE materials DB**, we chose to report E^{0-0} values.

High-throughput virtual screening library generation

VERDE materials DB relies on standardized high-throughput virtual screening (HTVS) libraries and an automated computational workflow. HTVS libraries are generated using an in-house algorithm that systematically links 20 spacer and 11 terminal groups shown in Scheme 2. These linking reactions are meant to resemble well-established cross-coupling reactions.[28]



Scheme 2. Combinatorial method used for generating high-throughput virtual screening libraries. Spacer groups are attached at user-defined substitution positions, then each spacer is combined with a terminal group.

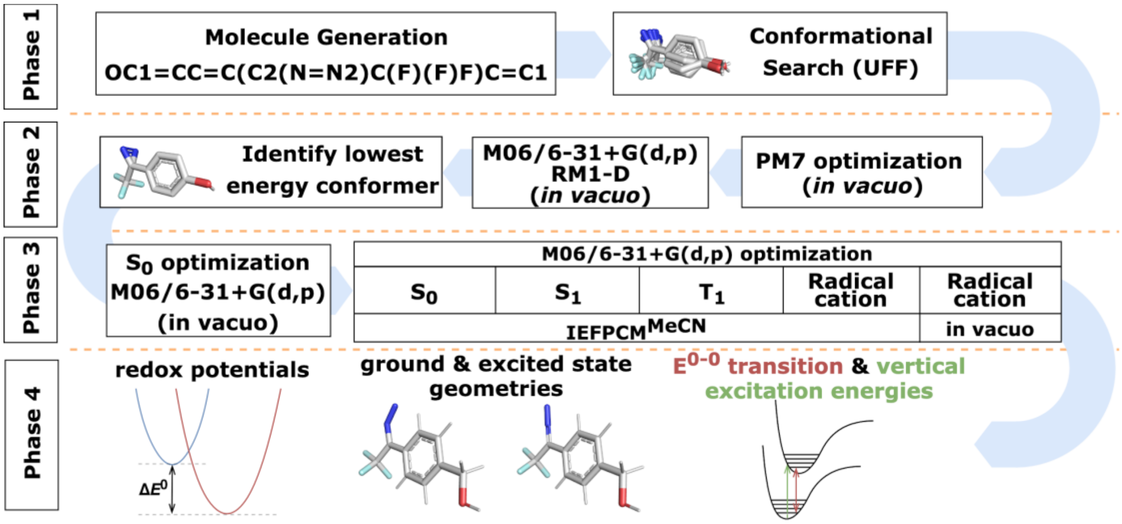
Generated molecules are then processed through the computational workflow illustrated in Scheme 3. The workflow is composed of four phases (computational details are elaborated in the computational methods section).

Phase 1: The workflow uses RDKit[29] to generate 3-D coordinates from the SMILES[30] string followed by a low-mode conformational search that produces up to four low-lying conformers minimized with the Universal Force Field.[31]

Phase 2: Each conformer in this ensemble is refined with two sequential semi-empirical optimization calculations: PM7[32] followed by RM1-D,[33] which includes the empirical D3-dispersion correction.[34] Our group has shown that RM1-D produces geometries that are remarkably close to DFT-optimized (ωB97XD/jun-cc-pvdz) structures. We then perform M06/6-31+G(d,p)[35-37] single point energy calculations on each of these optimized structures to determine the lowest-energy conformer.

Phase 3 and 4: The lowest-energy structure is subjected to an M06/6-31+G(d,p) optimization (with IEFPCM^{MeCN} to account for bulk solvent effects)[38] and frequency calculation to confirm the stationary point as true minimum on the ground- and excited-states (S_0, S_1, and T_1). This provides optimized geometries and $E^{0\text{-}0}$ values. The optimized structure and energies of the molecular radical cations afford the redox potentials of each molecule in the database.

Scheme 3. Illustration of the automated computational workflow used to run calculations for **VERDE materials DB**.



VERDE materials DB and MDF Integration

VERDE materials DB leverages MDF-operated data services, MDF Connect, Publish, and Discover to allow for streamlined access to, and discovery of, the data by researchers.[25] **MDF Publish** is a decentralized dataset repository, that allows a user to publish a dataset to any Globus endpoint,[39, 40] in the process creating a permanent identifier (e.g., DOI) for the dataset, and following a defined user-driven dataset curation flow to help ensure data quality. **MDF Discover** is an access-controlled, cloud-hosted search index with supportive Python software tools that support data search and facilitate data retrieval. **MDF Connect** is a service that supports the flow of data provided by a user from many storage locations to many services in the scientific data ecosystem. MDF Connect supports three key actions: 1) <u>submission</u> via user requests, made by script or web interface, triggers MDF Connect to collect the data from common storage locations including Google Drive, Box, or a Globus endpoint; 2) <u>enrichment</u> of collected data through extraction of general and domain-specific metadata (e.g., molecular information from output files, .xyz, .mol and other common chemistry data formats), combination of extracted and user-provided metadata into MDF metadata records, and transformation of dataset contents (e.g., from proprietary to open formats); and 3) <u>dispatch</u> of data to MDF Publish, metadata to MDF Discover, and combinations of data and metadata to other community data services (*e.g.*, NIST Materials Resource Registry, Citrine) selected by the user.

In the case of **VERDE materials DB**, data generated through high-throughput computations are submitted to MDF Connect via an automated Python script as it becomes available. Following submission, MDF Connect extracts important metadata describing the molecule being studied (e.g., InChI and SMILES strings, molecular mass) as well as calculated properties (e.g., dipole moments, redox potentials, and 0-0 transition energies) from files included in the submission to improve data discoverability (for a full description of the extracted metadata, see SI). These metadata are dispatched to MDF Discover where it is loaded into a search index to facilitate discovery and usage and then dispatched to MDF Publish to create a dataset, mint a permanent identifier, and move the data to storage endpoints at Argonne National Laboratory's Petrel

Facility[41] and at the University of Illinois at Urbana-Champaign Blue Waters to hold the associated data files and metadata. Further, **VERDE Materials DB** is made available to other community services for example, the NIST Materials Resource Registry for dataset registration. All dataset contents can be accessed via REST API or with the MDF Forge Python client. Users may then discover and download the entire dataset contents or the results of individual or matching calculations using the MDF Discover API and Python client.

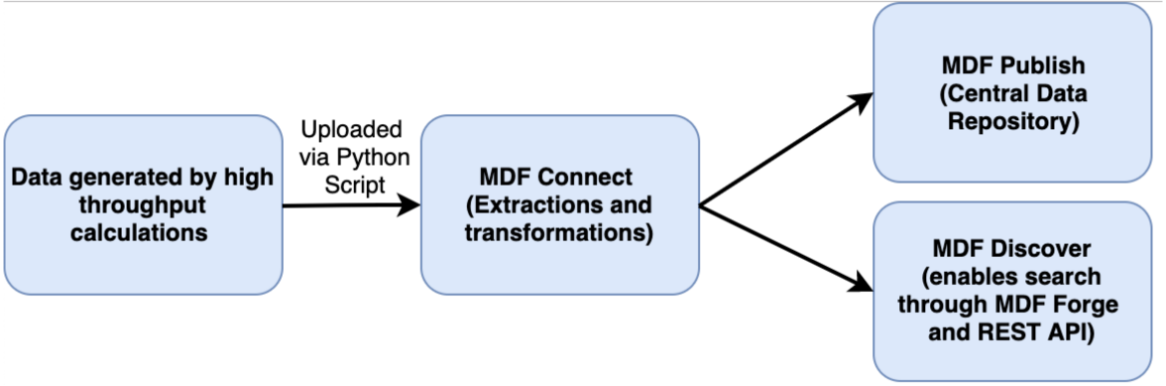


Figure 1. Data flow overview. Excited state data are submitted to MDF Connect. MDF Connect automatically extracts metadata (e.g., redox potentials, dipole moments, and 0-0 transition energies) from the submission, and the data and extracted metadata are dispatched to MDF Publish for long-term data storage, minting of a persistent identifier, and versioning, and to MDF Discover for loading into a search index to facilitate querying, aggregation, and data consumption.

Case-study: Screening of new photoaffinity labels

Photoaffinity labeling (PAL) is a technique used to identify the binding site of a protein through the regulated, covalent addition of a photoactive moiety to the protein of interest.[42] A ligand functionalized with a photoactive group can be irradiated once the ligand binds to the target protein, covalently binding the ligand to the protein. Further spectrometric analysis can be used to elucidate the location of binding. Knowledge of the binding site can guide the rational design of compounds that bind more strongly and specifically to the target protein.[42-44] Diazirines are a commonly used class of photoaffinity labels (PALs) and aryldiazirines, in particular, are known for their chemical and thermal stability, especially compared to other photoaffinity labels such as azides and benzophenones.[42, 45] Further, diazirines are one of the smallest photoreactive groups (PGs) used in PALs and therefore result in PALs that better mimic the ligand than do larger PGs.[45] Diazirines which exhibit greater photostability are desired to increase ambient light stability and increase the fidelity and specificity of PAL. Kumar *et al*. has determined the photostabilities of the 3-trifluoromethyl-3-aryldiazirines, which are summarized in Figure 2.[46]

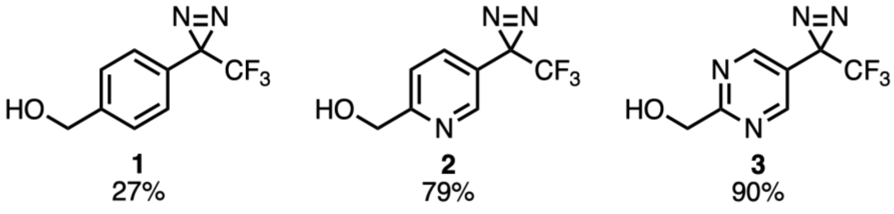
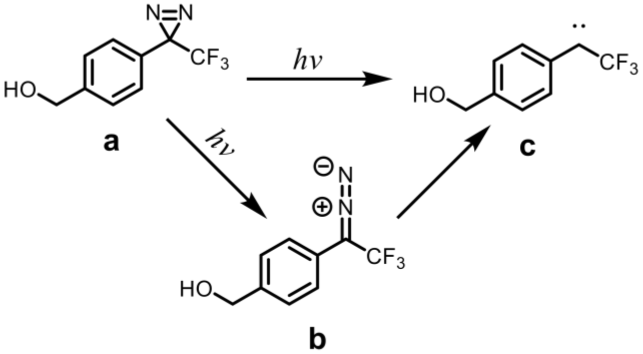


Figure 2. Percentage of 3-trifluoromethyl-3-aryldiazirines remaining after 31 days of ambient light exposure.

The 0-0 transition energy (E^{0-0}) is defined in Eq. 1 as the difference in energy between the S_0 and S_1 states minus the difference in zero-point vibrational energy ($\Delta ZPVE$) between the two states. This value empirically corresponds to the midpoint between the λ^{max} of the emission and absorption spectra.

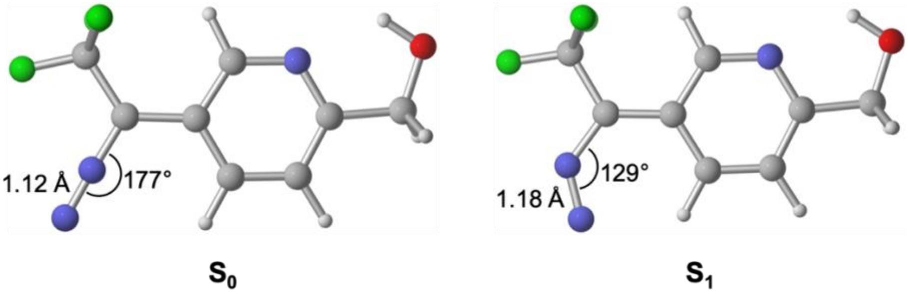
$$E^{0-0} = E_{S_1} - E_{S_0} - \Delta ZPVE \qquad \text{Eq. (1)}$$

E^{0-0} values were computed for the aryldiazirines in Figure 2. The computed E^{0-0} values are 2.85, 2.91, and 3.01 eV for compounds **1**, **2**, and **3** respectively. Larger E^{0-0} energies correspond with increasing stability along the series shown in Figure 2. This trend appears to correlate with the electron-withdrawing nature of the substituents, with more strongly electron-withdrawing substituents resulting in greater E^{0-0} values. Our groups were also interested in understanding the concerted or stepwise nature of the photochemical diazirine ring-opening mechanism (Scheme 4), which is largely outside the scope of this manuscript.



Scheme 4. Representative concerted and stepwise mechanisms of carbene formation in a prototypical diazirine system. (a) The initial diazirine, (b) the diazo intermediate in the stepwise mechanism, and (c) the carbene product.

However, the diazo intermediate (Figure 3), had a substantially different geometry in the S_1 state than in its ground state because of $\pi_{NN}\pi^*_{NN}$ transition that rehybridizes the central nitrogen of the diazo intermediate. This photoexcited geometry is remarkably close–and thus pre-distorted–towards an adjacent conical intersection.



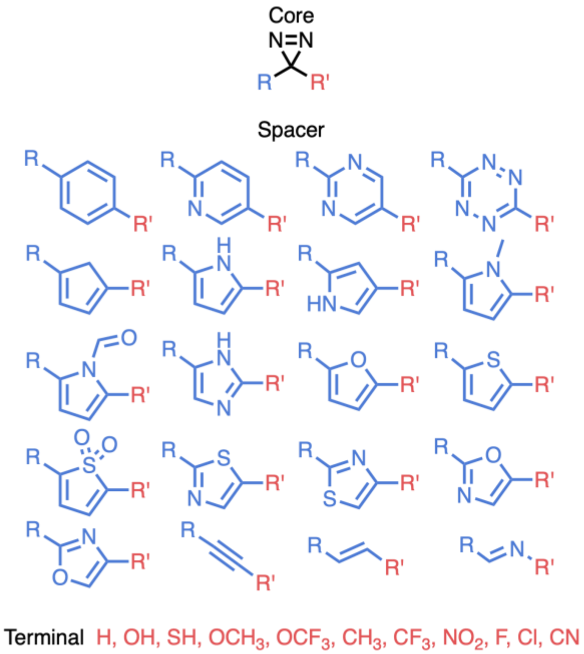
S_0

S_1

Figure 3. Geometries of diazo intermediate in the ground (S_0) and excited (S_1) states. The length of the N=N bond and the C=N=N bond angle are labeled.

The mechanistic study in on-going between our groups and will be published in due course. After consulting with Manetsch and co-workers, we jointly designed a virtual library of 206 diazirines (following a modified combinatorial method shown in Scheme 5) that were subjected to the workflow in Scheme 3.

Scheme 5. Combinatorial method for generating HTVS library of diazirines



These 206 diazirines had E^{0-0} values ranging from 1.60 to 3.87 eV. We identified 41 candidates that met the E^{0-0} values > 3.00eV, informed by the high photostability of compound **3** (Figure 2). Figure 4 shows a subset of these top candidates (See SI for full list).

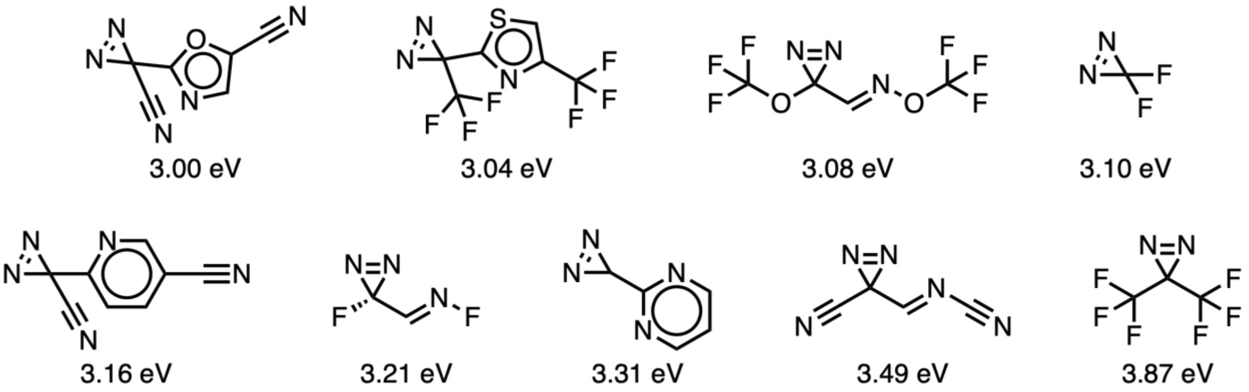


Figure 4. Theoretical diazirines with E^{0-0} > 3.00 eV. Geometry optimizations and frequency calculations were performed with M06/6-31+G(d,p) and E^{0-0} values were computed.

The greatest E^{0-0} correspond to those diazirines with electron-withdrawing substituents. In these diazirines, the HOMOs are strongly stabilized by the electron-withdrawing group and the LUMOs are relatively unperturbed, leading to larger excitation energies, manifested as higher E^{0-0} transition energies. Experimental verification of these results is currently underway in our groups; we anticipate that the diazirines in Figure 4 will be at least as stable as **3**. Solar and fluorescent light have vanishingly few photons in UV-range; those diazirines requiring relatively high-energy photons will be slower to react under these standard illumination conditions.

Conclusions

Our pyMolGen code and high throughput virtual screening workflow has been used to determine the ground- and excited-state structures (S_0, S_1, T_1) for 1,282 organic π-conjugated organic molecules. We established the **VERDE materials DB** as the first database to feature these optimized geometries and corresponding properties, including redox potentials, dipole moments, excitation energies, redox potentials, and 0-0 transition energies. It is hosted on the Materials Data Facility in a form conducive to consumption by researchers, and is continually growing through internal projects and collaborations. Data gathered from high-throughput virtual screening (HTVS) of diazirine derivatives for photoaffinity labeling showcases the utility of computed excited structures and properties. Ongoing HTVS projects are examining dibenzoperylene and anthraquinone derivatives for organic photoredox catalysis and singlet fission materials, respectively.We are working to make **VERDE materials DB** even more interactive by including features for users to upload structures to be computed with our resources.

Computational Methods

We developed a Python-based code that generates molecular SMILES[30] strings based on a π-conjugated core moiety with substituent sites informed by literature and commercial precedent. These SMILES strings are organized in a virtual screening library to begin the automatic computational workflow. We employ the RDKit[29] Python library to generate four conformers, which undergo structural relaxation with the Universal Force Field. Each conformer is subjected to the following series of calculations: (1) PM7[32] optimization, (2) RM1-D optimization (using DFT-D3 dispersion correction),[33, 34] and (3) single-point DFT calculation. The lowest energy conformer, determined based on the aforementioned single-point DFT calculation, undergoes the following series of DFT optimizations and frequency calculations (4) S_0 *in vacuo*, (5) S_0 in $\text{IEFPCM}^{\text{MeCN}}$, (6) S_1 in $\text{IEFPCM}^{\text{MeCN}}$, (7) T_1 in $\text{IEFPCM}^{\text{MeCN}}$, (8) radical cation *in vacuo*, (9) radical cation in $\text{IEFPCM}^{\text{MeCN}}$.[38] All DFT calculations are performed by using M06/6-31+G(d,p).[35-37] All calculations are performed with the default settings provided by the Gaussian 16 software package,[47] with the exception of the RM1-D optimization which is performed by using GAMESS version 2018 R1.[48, 49] 0-0 transition energies are derived from these calculations based on a method described by Jacquemin *et al.*[26] Redox and ionization potentials are computed as described by Fu *et al.*[27] Excited state redox potentials are computed as described by Romero *et al.*[10]

Acknowledgements

Biruk Abreha acknowledges the Northeastern University Office of Undergraduate Research and Fellowship for the Summer Scholars Independent Research Fellowship. We would also like to thank Northeastern University Research Computing and the Chemistry & Chemical Biology department for start-up funding. This work was supported in part under financial assistance award 70NANB19H005 from the U.S. Department of Commerce, National Institute of Standards and Technology as part of the Center for Hierarchical Materials Design (CHiMaD), and by U.S. Department of Energy Contract DE-AC02-06CH11357.

Bibliography

1. IEA *Key World Energy Statistics 2018*; Paris, 2018; p 16.
2. Kabir, E.; Kumar, P.; Kumar, S.; Adelodun, A. A.; Kim, K.-H., Solar energy: Potential and future prospects. *Renewable and Sustainable Energy Reviews* **2018,** *82*, 894-900.
3. Green, M. A.; Dunlop, E. D.; Levi, D. H.; Hohl-Ebinger, J.; Yoshita, M.; Ho-Baillie, A. W. Y., Solar cell efficiency tables (version 54). *Progress in Photovoltaics: Research and Applications* **2019,** *27* (7), 565-575.
4. Cui, Y.; Yao, H.; Zhang, J.; Zhang, T.; Wang, Y.; Hong, L.; Xian, K.; Xu, B.; Zhang, S.; Peng, J.; Wei, Z.; Gao, F.; Hou, J., Over 16% efficiency organic photovoltaic cells enabled by a chlorinated acceptor with increased open-circuit voltages. *Nat Commun* **2019,** *10* (1), 2515.
5. Huskinson, B.; Marshak, M. P.; Suh, C.; Er, S.; Gerhardt, M. R.; Galvin, C. J.; Chen, X.; Aspuru-Guzik, A.; Gordon, R. G.; Aziz, M. J., A metal-free organic-inorganic aqueous flow battery. *Nature* **2014,** *505* (7482), 195-8.
6. Ding, Y.; Zhang, C.; Zhang, L.; Zhou, Y.; Yu, G., Molecular engineering of organic electroactive materials for redox flow batteries. *Chem Soc Rev* **2018,** *47* (1), 69-103.
7. Sampaio, P. G. V.; González, M. O. A., Photovoltaic solar energy: Conceptual framework. *Renewable and Sustainable Energy Reviews* **2017,** *74*, 590-601.
8. Babayigit, A.; Ethirajan, A.; Muller, M.; Conings, B., Toxicity of organometal halide perovskite solar cells. *Nat Mater* **2016,** *15* (3), 247-51.
9. Cheng, P.; Li, G.; Zhan, X.; Yang, Y., Next-generation organic photovoltaics based on non-fullerene acceptors. *Nat. Photon.* **2018,** *12* (3), 131-142.
10. Romero, N. A.; Nicewicz, D. A., Organic Photoredox Catalysis. *Chem Rev* **2016,** *116* (17), 10075-166.
11. Smith, E.; Collins, I., Photoaffinity labeling in target- and binding-site identification. *Future Med. Chem.* **2015,** *7* (2), 159–83.
12. Smith, M. B.; Michl, J., Singlet fission. *Chem Rev* **2010,** *110* (11), 6891-936.
13. Polishchuk, P. G.; Madzhidov, T. I.; Varnek, A., Estimation of the size of drug-like chemical space based on GDB-17 data. *J Comput Aided Mol Des* **2013,** *27* (8), 675-9.
14. Hachmann, J.; Olivares-Amaya, R.; Atahan-Evrenk, S.; Amador-Bedolla, C.; Sánchez-Carrera, R. S.; Gold-Parker, A.; Vogt, L.; Brockway, A. M.; Aspuru-Guzik, A., The Harvard Clean Energy Project: Large-Scale Computational Screening and Design of Organic Photovoltaics on the World Community Grid. *The Journal of Physical Chemistry Letters* **2011,** *2* (17), 2241-2251.
15. Lopez, S. A.; Pyzer-Knapp, E. O.; Simm, G. N.; Lutzow, T.; Li, K.; Seress, L. R.; Hachmann, J.; Aspuru-Guzik, A., The Harvard organic photovoltaic dataset. *Sci Data* **2016,** *3*, 160086.
16. Scharber, M. C.; Mühlbacher, D.; Koppe, M.; Denk, P.; Waldauf, C.; Heeger, A. J.; Brabec, C. J., Design Rules for Donors in Bulk-Heterojunction Solar Cells—Towards 10 % Energy-Conversion Efficiency. *Advanced Materials* **2006,** *18* (6), 789-794.
17. Blum, L. C.; Reymond, J.-L., 970 Million Druglike Small Molecules for Virtual Screening in the Chemical Universe Database GDB-13. *J. Am. Chem. Soc.* **2009,** *131*, 8732–8733.
18. Rupp, M.; Tkatchenko, A.; Muller, K. R.; von Lilienfeld, O. A., Fast and accurate modeling of molecular atomization energies with machine learning. *Phys Rev Lett* **2012,** *108* (5), 058301.

19. Montavon, G.; Rupp, M.; Gobre, V.; Vazquez-Mayagoitia, A.; Hansen, K.; Tkatchenko, A.; Muller, K. R.; von Lilienfeld, O. A., Machine Learning of Molecular Electronic Properties in Chemical Compound Space. *New Journal of Physics* **2013,** *15*.
20. Ramakrishnan, R.; Dral, P. O.; Rupp, M.; von Lilienfeld, O. A., Quantum chemistry structures and properties of 134 kilo molecules. *Sci Data* **2014,** *1*, 140022.
21. Gilmer, J.; Schoenholz, S. S.; Riley, P. F.; Vinyals, O.; Dahl, G. E., Neural Message Passing for Quantum Chemistry. In *Proceedings of the 34th International Conference on Machine Learning*, Sydney, NSW, Australia, 2017; Vol. 70, pp 1263–72.
22. Gomez-Bombarelli, R.; Wei, J. N.; Duvenaud, D.; Hernandez-Lobato, J. M.; Sanchez-Lengeling, B.; Sheberla, D.; Aguilera-Iparraguirre, J.; Hirzel, T. D.; Adams, R. P.; Aspuru-Guzik, A., Automatic Chemical Design Using a Data-Driven Continuous Representation of Molecules. *ACS Cent Sci* **2018,** *4* (2), 268-276.
23. Smith, J. S.; Isayev, O.; Roitberg, A. E., ANI-1: an extensible neural network potential with DFT accuracy at force field computational cost. *Chem Sci* **2017,** *8* (4), 3192-3203.
24. Gomez-Bombarelli, R.; Aguilera-Iparraguirre, J.; Hirzel, T. D.; Duvenaud, D.; Maclaurin, D.; Blood-Forsythe, M. A.; Chae, H. S.; Einzinger, M.; Ha, D. G.; Wu, T.; Markopoulos, G.; Jeon, S.; Kang, H.; Miyazaki, H.; Numata, M.; Kim, S.; Huang, W.; Hong, S. I.; Baldo, M.; Adams, R. P.; Aspuru-Guzik, A., Design of efficient molecular organic light-emitting diodes by a high-throughput virtual screening and experimental approach. *Nat Mater* **2016,** *15* (10), 1120-7.
25. Blaiszik, B.; Ward, L.; Schwarting, M.; Gaff, J.; Chard, R.; Pike, D.; Chard, K.; Foster, I., A Data Ecosystem to Support Machine Learning in Materials Science. arXiv:1904.10423 [cond-mat.mtrl-sci], 2019.
26. Jacquemin, D.; Planchat, A.; Adamo, C.; Mennucci, B., TD-DFT Assessment of Functionals for Optical 0-0 Transitions in Solvated Dyes. *J Chem Theory Comput* **2012,** *8* (7), 2359-72.
27. Fu, Y.; Liu, L.; Yu, H.-Z.; Wang, Y.-M.; Guo, Q.-X., Quantum-Chemical Predictions of Absolute Standard Redox Potentials of Diverse Organic Molecules and Free Radicals in Acetonitrile. *J. Am. Chem. Soc.* **2004,** *127*, 7227–7234.
28. Ruiz-Castillo, P.; Buchwald, S. L., Applications of Palladium-Catalyzed C-N Cross-Coupling Reactions. *Chem Rev* **2016,** *116* (19), 12564-12649.
29. Landrum, G. *RDKit: Open-Source Cheminformatics and Machine Learning*.
30. Weininger, D., SMILES, a chemical language and information system. 1. Introduction to methodology and encoding rules. *Journal of Chemical Information and Modeling* **1988,** *28* (1), 31-36.
31. Rappé, A. K.; Casewit, C. J.; Colwell, K. S.; III, W. A. G.; Skiff, W. M., UFF, a Full Periodic Table Force Field for Molecular Mechanics and Molecular Dynamics Simulations. *J. Am. Chem. Soc.* **1992,** *114* (25), 10024–10035.
32. Frisch, M. J.; Throssel, K., Evaluation and Improvement of Semi-empirical methods I: PM7R8: A variant of PM7 with numerically stable hydrogen bonding corrections. *in prep*.
33. Rocha, G. B.; Freire, R. O.; Simas, A. M.; Stewart, J. J., RM1: a reparameterization of AM1 for H, C, N, O, P, S, F, Cl, Br, and I. *J Comput Chem* **2006,** *27* (10), 1101-11.
34. Grimme, S.; Antony, J.; Ehrlich, S.; Krieg, H., A consistent and accurate ab initio parametrization of density functional dispersion correction (DFT-D) for the 94 elements H-Pu. *J Chem Phys* **2010,** *132* (15), 154104.

35. Frisch, M. J.; Pople, J. A.; Binkley, J. S., Self-consistent molecular orbital methods 25. Supplementary functions for Gaussian basis sets. *The Journal of Chemical Physics* **1984,** *80* (7), 3265-3269.

36. Hehre, W. J.; Ditchfield, R.; Pople, J. A., Self—Consistent Molecular Orbital Methods. XII. Further Extensions of Gaussian—Type Basis Sets for Use in Molecular Orbital Studies of Organic Molecules. *The Journal of Chemical Physics* **1972,** *56* (5), 2257-2261.

37. Zhao, Y.; Truhlar, D. G., The M06 suite of density functionals for main group thermochemistry, thermochemical kinetics, noncovalent interactions, excited states, and transition elements: two new functionals and systematic testing of four M06-class functionals and 12 other functionals. *Theoretical Chemistry Accounts* **2007,** *120* (1-3), 215-241.

38. Tomasi, J.; Mennucci, B.; Cammi, R., Quantum Mechanical Continuum Solvation Models. *Chem Rev* **2005,** *105* (8).

39. Ananthakrishnan, R.; Blaiszik, B.; Chard, K.; Chard, R.; McCollam, B.; Pruyne, J.; Rosen, S.; Tuecke, S.; Foster, I., Globus Platform Services for Data Publication. In *Proceedings of the Practice and Experience on Advanced Research Computing - PEARC '18*, 2018; pp 1-7.

40. Chard, K.; Foster, I.; Tuecke, S., Globus: Research Data Management as Service and Platform. In *Proceedings of the Practice and Experience in Advanced Research Computing 2017 on Sustainability, Success and Impact - PEARC17*, 2017; pp 1–5.

41. Allcock, W. E.; Wagner, R.; Allen, B. S.; Ananthakrishnan, R.; Blaiszik, B.; Chard, K.; Chard, R.; Foster, I.; Lacinski, L.; Papka, M. E., Petrel: A Programmatically Accessible Research Data Service. In *Proceedings of the Practice and Experience in Advanced Research Computing on Rise of the Machines (learning) - PEARC '19*, 2019; pp 1-7.

42. Smith, E.; Collins, I., Photoaffinity labeling in target- and binding-site identification. *Future Med Chem* **2015,** *7* (2), 159-83.

43. Hamouda, A. K.; Stewart, D. S.; Chiara, D. C.; Savechenkov, P. Y.; Bruzik, K. S.; Cohen, J. B., Identifying barbiturate binding sites in a nicotinic acetylcholine receptor with [3H]allyl m-trifluoromethyldiazirine mephobarbital, a photoreactive barbiturate. *Mol Pharmacol* **2014,** *85* (5), 735-46.

44. McKernan, R. M.; Farrar, S.; Collins, I.; Emms, F.; Asuni, A.; Quirk, K.; Broughton, H., Photoaffinity Labeling of the Benzodiazepine Binding Site of α1β3γ2 γ-Aminobutyric Acid$_A$ Receptors with Flunitrazepam Identifies a Subset of Ligands that Interact Directly with His102 of the α Subunit and Predicts Orientation of These within the Benzodiazepine Pharmacophore. *Mol Pharmacol* **1998,** *54* (1), 33–43.

45. Dubinsky, L.; Krom, B. P.; Meijler, M. M., Diazirine based photoaffinity labeling. *Bioorg Med Chem* **2012,** *20* (2), 554-70.

46. Kumar, A. B.; Tipton, J. D.; Manetsch, R., 3-Trifluoromethyl-3-aryldiazirine photolabels with enhanced ambient light stability. *Chem Commun (Camb)* **2016,** *52* (13), 2729-32.

47. Frisch, M. J. T., G. W.; Schlegel, H. B.; Scuseria, G. E.; Robb, M. A.; Cheeseman, J. R.; Scalmani, G.; Barone, V.; Petersson, G. A.; Nakatsuji, H.; Li, X.; Caricato, M.; Marenich, A. V.; Bloino, J.; Janesko, B. G.; Gomperts, R.; Mennucci, B.; Hratchian, H. P.; Ortiz, J. V.; Izmaylov, A. F.; Sonnenberg, J. L.; Williams-Young, D.; Ding, F.; Lipparini, F.; Egidi, F.; Goings, J.; Peng, B.; Petrone, A.; Henderson, T.; Ranasinghe, D.; Zakrzewski, V. G.; Gao, J.; Rega, N.; Zheng, G.; Liang, W.; Hada, M.; Ehara, M.; Toyota, K.; Fukuda, R.; Hasegawa, J.; Ishida, M.; Nakajima, T.; Honda, Y.; Kitao, O.; Nakai, H.; Vreven, T.; Throssell, K.; Montgomery, J. A., Jr.; Peralta, J. E.; Ogliaro, F.; Bearpark, M. J.; Heyd, J. J.; Brothers, E. N.; Kudin, K. N.; Staroverov, V. N.; Keith, T. A.; Kobayashi, R.; Normand, J.; Raghavachari, K.; Rendell, A. P.; Burant, J. C.;

Iyengar, S. S.; Tomasi, J.; Cossi, M.; Millam, J. M.; Klene, M.; Adamo, C.; Cammi, R.; Ochterski, J. W.; Martin, R. L.; Morokuma, K.; Farkas, O.; Foresman, J. B.; Fox, D. J. *Gaussian 16*, Revision A.03; Gaussian, Inc.: Wallingford, CT, 2016.
48. Gordon, M. S.; Schmidt, M. W., Advances in electronic structure theory: GAMESS a decade later. In *Theory and Applications of Computational Chemistry*, Dykstra, C. E.; Frenking, G.; Kim, K. S.; Scuseria, G. E., Eds. Elsevier: 2005; pp 1167–1189.
49. Schmidt, M. W.; Baldridge, K. K.; Boatz, J. A.; Elbert, S. T.; Gordon, M. S.; Jensen, J. H.; Koseki, S.; Matsunaga, N.; Nguyen, K. A.; Su, S.; Windus, T. L.; Dupuis, M.; Montgomery, J. A., General atomic and molecular electronic structure system. *J Comput Chem* **1993,** *14* (11), 1363.