PAL 2.0: A Physics-Driven Bayesian Optimization Framework for Material Discovery

Maitreyee Sharma Priyadarshini*,[1] Oluwaseun Romiluyi*,[1] Yiran Wang,[1] Kumar Miskin,[2] Connor Ganley,[1] and Paulette Clancy[1, a)]

[1)] *Department of Chemical and Biomolecular Engineering, Johns Hopkins University,
3400 North Charles Street, Baltimore, 21218, Maryland, USA*

[2)] *Department of Materials Science and Engineering, Johns Hopkins University,
3400 North Charles Street, Baltimore, 21218, Maryland, USA*

(Dated: November 13, 2023)

[a)] Electronic mail: pclancy3@jhu.edu

CONTENTS

S1. NOMENCLATURE

Given the considerably variable nomenclature used across the literature to describe computational material discovery and machine learning, we describe below the nomenclature used throughout this paper in order to provide clarity to the reader.

First, the building blocks of the materials being studied are referred to as *design features.* For example, in the case of a perovskite crystal, characterized by the chemical formula, ABX_3, the design features would entail choices of the A-site cation, B-site cation and X-site anion. The different choices for each design feature are referred to as *design choices.* In the perovskite crystal example, the X-site anion can be Cl^-, I^- or Br^-, and these choices are called design choices. The A-site cation could be methylammonium (MA), formamadinium (FA) of caesium (Cs), for instance, and so on.

Additionally, each design feature can either be represented as a one-hot-encoded (OHE) vector[1] or by various physical properties describing the design feature. For example, the X-site anion can be described using its electronegativity, ionic radius, *etc.* We refer to these physical properties as *descriptors.*

The set of all descriptors provided by the user for all design features is called the *property basket.* The above nomenclature describes the material domain.

We now describe the nomenclature used to describe the method developed in this work. The material domain on which the search is conducted is represented by *Gaussian Process Regression (GPR)*[2] models. The choice of different prior mean choices are represented by *GP-Prior Choice*, for example if we use a linear function as the prior mean, the model is written as GP-L. The GPR models are constructed using descriptors for design features.

Lastly, the nomenclature referring to the optimization algorithm includes *Bayesian Optimization (BayesOpt)*[3], the *surrogate model* for BayesOpt given by the GPR models and the *acquisition function*, used to refer to the function that gives a recommendation for the next point to consider.

S2. STATE-OF-THE-ART METHODS

In this section, we briefly describe state-of-the-art methods that are currently used for material discovery.

Random search: A straightforward, if brute-force, approach to finding the optimal material relies on randomly and exhaustively exploring the material space. On average, random searches will require exploration of 50% (or more) of the material space in order to find the optimal candidates. Therefore, this search strategy is applicable only when the size of the space being explored is small and the cost to explore the space is relatively low.

Bayesian optimization: Various extensions of the Bayesian optimization algorithm have been developed that are also applicable to material discovery. One such framework is called SMAC (Sequential Model-based Algorithm Configuration)[4]. SMAC is a versatile Bayesian optimization package for automated algorithm configuration and hyperparameter optimization. SMAC offers many default optimization pipelines, and packages them into different interfaces called facades. In this work, we mainly explore SMAC's *Black-Box Facade*, which contains a Gaussian Process as its surrogate model and Expected Improvement as its acquisition function, and its *Hyperparameter Optimization Facade* which includes Random Forest and Log Expected Improvement as its surrogate model and acquisition function, respectively.

Likewise, Hyperopt[5] is another popular Python library used for hyperparameter optimization for machine learning tools. It employs a combination of randomized search and tree-structured Parzen estimators (TPE) to efficiently explore the hyperparameter space and find optimal configurations. Like SMAC, it also uses Bayesian optimization to intelligently select hyperparameter settings to evaluate based on results from previous trials. The results we present in this paper have been benchmarked against these methods; see Section 2.

S3. DOPED P-TYPE ORGANIC SEMICONDUCTING POLYMERS

A. Data Generation

For the density functional theory (DFT) calculations in this data set, the effects of solvent screening were approximated through a conductor-like polarizable continuum model (CPCM).[6] The geometry optimizations and single-point calculations were carried out using the ORCA software[7] with a B97-D3 functional and def2-TZVP basis set, per the recommendations from Goerigk and Grimme for organic molecules.[8] For greater detail about the raster grid-based approach employed in the binding enthalpy calculations, we direct the reader to the Supplemental Information of Mukhopadhyaya et al.[9]

B. Description of Properties in the Property Basket

DPP	No. of diketopyrrolopyrrole (DPP) structures in the polymer repeat unit
EDOT	No. of ethylenedioxythiophene (EDOT) structures in the polymer repeat unit
TT	No. of thienothiophene (TT) structures in the polymer repeat unit
Benzene Presence	No. of benzene rings in the polymer repeat unit
Num Arms	No. of chains emanating radially from the repeat unit's center
EN DIFF	Absolute value of the difference between the most and least electronegative atoms in the polymer repeat unit (Pauling units)
HOMO	DFT-calculated HOMO (eV)
LUMO	DFT-calculated LUMO (eV)
HL	Absolute value of the difference between the repeat unit's HOMO and LUMO (eV)
MW	Molecular weight of repeat unit (g/mol)
Molecular Weight	Molecular weight of dopant (g/mol)
HOMO	DFT-calculated HOMO (eV)
LUMO	DFT-calculated LUMO (eV)
HOMO-LUMO gap (eV)	Absolute value of the difference between the repeat unit's HOMO and LUMO (eV)
—EN DIFF—	Absolute value of the difference between the most and least electronegative atoms in the polymer repeat unit (Pauling units)
Dielectric	Dielectric constant applied in ORCA's CPCM ”implicit solvation” implementation
Refractive Index	Refractive index applied in ORCA's CPCM ”implicit solvation” implementation
Molecular Weight	Molecular weight of solvent molecule (g/mol)
TPSA	Topological polar surface area ($\mathring{A}^2$)
Complexity	Molecular complexity provides an estimate of the synthetic effort
DN	Gutmann Donor Number (kcal/mol)
AN	Acceptor Number

Table S1: Description of properties in the property basket of the doped p-type organic semiconductor dataset shown in Fig. 3.

C. Properties Selected by XGBoost

DPP	No. of diketopyrrolopyrrole (DPP) structures in the polymer repeat unit
EDOT	No. of ethylenedioxythiophene (EDOT) structures in the polymer repeat unit
TT	No. of thienothiophene (TT) structures in the polymer repeat unit
Benzene Presence	No. of benzene rings in the polymer repeat unit
HOMO	DFT-calculated HOMO (eV)
Molecular Weight	Molecular weight of dopant (g/mol)
LUMO	DFT-calculated LUMO (eV)
Dielectric	Dielectric constant applied in ORCA's CPCM ”implicit solvation” implementation
Refractive Index	Refractive index applied in ORCA's CPCM ”implicit solvation” implementation
TPSA	Topological polar surface area ($\mathring{A}^2$)
Complexity	Molecular complexity provides an estimate of the synthetic effort

Table S2: Description of properties selected by XGBoost that are used as the input variable to train the Neural Network mean function of the Gaussian Process model. These are properties selected for the doped p-type organic semiconductor dataset shown in Fig. 3.

S4. METAL HALIDE PEROVSKITES BANDGAP DATASET

A_ion_rad	A-site atom/molecule Ionic Radius (Angstroms)
A_den	A-site atom/molecule density (g/cm^3)
A_at_wt	A-site atomic/molecular weight (u)
A_IE	A-site atom/molecule Ionization Energy (kJ/mol)
B_ion_rad	B-site weighted average of atomic ionic radius (Angstroms)
B_den	B-site weighted average atom/molecule density (g/cm^3)
B_at_wt	B-site weighted average atomic/molecular weight (u)
B_EA	B-site weighted average atom/molecule Electron Affinity (kJ/mol)
B_IE	B-site weighted average atom/molecule Ionization Energy (kJ/mol)
B_EN	B-site weighted average atom/molecule Electronegativity
X_ion_rad	X-site of atomic ionic radius (kJ/mol)
X_den	X-site atomic density (g/cm^3)

Table S3: Description of properties selected by XGBoost that are used as the input variable to train the Neural Network mean function of the Gaussian Process model. These are properties selected for the metal halide perovskite bandgap dataset shown in Fig. 4. This dataset is taken from Ref. 10.

S5. METAL HALIDE PEROVSKITE AND SOLVENT PAIRS DATASET FOR SOLUTION-PROCESSED THIN FILM

Halide1-Electro	Halide 1 electronegativity
Halide2-Electro	Halide 2 electronegativity
Halide3-Electro	Halide 3 electronegativity
Cation-Radius	A-site atom/molecule Ionic Radius (Angstroms)
Cation-Enthalpy	A-site ion-DMF solvent binding enthalpy (kcal/mol)[11]
Solvent-DN	Gutmann Donor Number (kcal/mol)
Solvent-LCA	Lithium Cation Affinity (kcal/mol)
Solvent-AN	Acceptor Number
Solvent-DPM	Dipole Moment (Debye)
Solvent-Dielectric	Dielectric constant
Solvent-Density	Density (g/cm^3)
Solvent-MV	Molar volume (g/mol)

Table S4: Description of properties selected by XGBoost that are used as the input variable to train the Neural Network mean function of the Gaussian Process model. These are properties selected for the metal halide perovskite and solvent pair dataset shown in Fig. 5. This dataset is taken from Ref. 1.

S6. STRESS TESTS FOR THE PAL 2.0 METHOD

In this section, we provide two stress tests we did to evaluate the performance of the PAL 2.0 and understand its limitation.

For the first stress test, we evaluate the performance of the method with varying amounts of initial data provided for training the surrogate models before running Bayesian Optimization. This stress test is conducted on the metal halide perovskite bandgap dataset consisting of 244 materials.[10] We assess the BayesOpt performance for four initial data fractions used for training of the GP models before running BayesOpt: 5%, 10%, 25% and 50%, Fig. S1. It is observed that the GP-NN surrogate model created in PAL 2.0 outperforms the GP-0 model in all instances. However, the limitation of the GP-NN model is that it needs some amount of initial data to train the Neural Network prior mean function. On the other hand, the GP-0 model can be used without any training data.

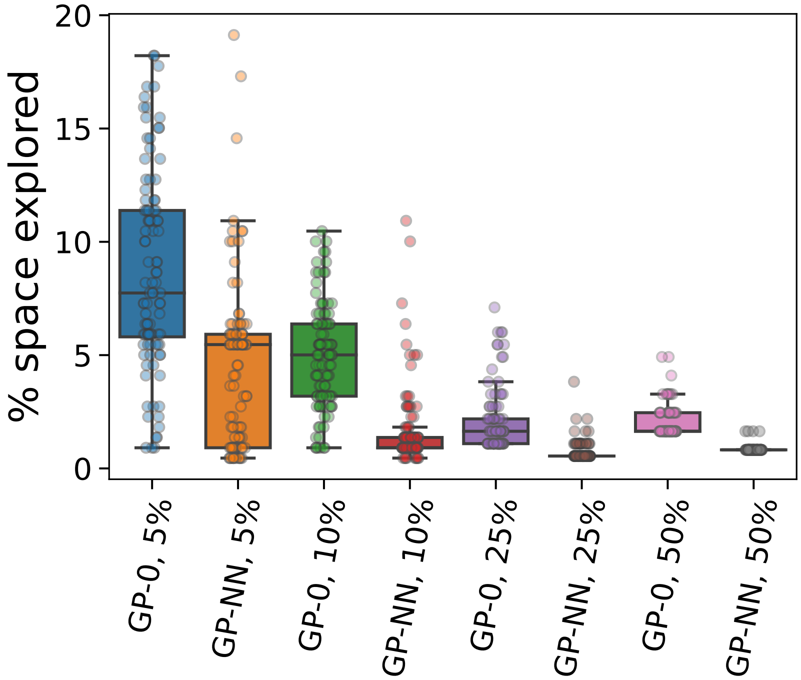

Figure S1: Stress test for PAL 2.0 for four initial data fractions used to train two different GP models (GP-0 and GP-NN, essentially with and without the use of a neural network step) before running BayesOpt: 5%, 10%, 25% and 50%. GP-NN outperforms GP-0 for all data fractions and also needs to explore less of the space. Even with only 5% of the data, the GP-NN (orange) has to explore less than 5% of the space. These results are shown on the metal halide perovskite band gap dataset.[10]

For the second stress test, we evaluate the performance of the PAL 2.0 framework on a covalent organic framework (COF) dataset that consists of 69,840 2D and 3D COFs.[12] This dataset replicates some real-world scenarios wherein the search space to explore is very large but the data available for training is limited. In this stress test, we used an initial training dataset of 68 points, *i.e.,* 0.1% of the total search space. The optimization target here is the deliverable capacity (v STP/v) of the COF structure. The results obtained are compared with the Bayesian Optimization results from a recent work by Deshwal *et al.*[13] In Fig. S2, we plot the average number of COFs that are evaluated before we find the optimal COF structure. The average reported for each COF tried is taken over 100 repetitions of the Bayesian Optimization with different initial training datasets that are chosen randomly. We observed in Fig. S2 that the GP-NN model from PAL 2.0 outperforms the GP-0 model and, indeed, the result by Deshwal *et al.*[13] by finding the optimal COF within just 4 COF evaluations. We attribute this superior performance of the GP-NN model to the highly predictive and accurate NN prior mean function. Since the NN prior mean is predictive, the GP-NN model already has an accurate representation of the optimization landscape. As a result, this model, in conjunction with the acquisition function, is able to rapidly point to the optimal location within 4 iterations of the Bayesian Optimization code. This testcase shows the validity of, and high performance of, the novel surrogate model that we have constructed in this paper.

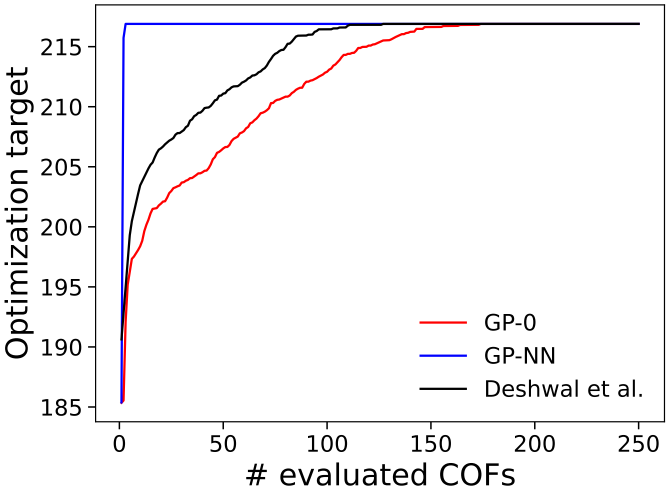

Figure S2: Convergence plot showing the average number of Covalent Organic Frameworks (COFs) evaluated before finding the most optimal COF for methane storage applications. The dataset consists of 69840 2D and 3D COFs.[12] The optimization target here is the deliverable capacity (v STP/v) of the COF structure. An initial training dataset of 68 points, i.e. 0.1% of the total search space, is used to train the priors of the Guassian Process models. The results obtained are compared with the Bayesian Optimization results from a recent work by Deshwal et al.[13]

REFERENCES

[1] H. C. Herbol, W. Hu, P. Frazier, P. Clancy and M. Poloczek, *npj Computational Materials*, 2018, **4**, 51.

[2] C. E. Rasmussen, in *Gaussian Processes in Machine Learning*, ed. O. Bousquet, U. von Luxburg and G. Rätsch, Springer Berlin Heidelberg, Berlin, Heidelberg, 2004, pp. 63–71.

[3] J. Snoek, H. Larochelle and R. P. Adams, Advances in Neural Information Processing Systems, 2012.

[4] M. Lindauer, K. Eggensperger, M. Feurer, A. Biedenkapp, D. Deng, C. Benjamins, T. Ruhkopf, R. Sass and F. Hutter, *Journal of Machine Learning Research*, 2022, **23**, 1–9.

[5] J. Bergstra, D. Yamins and D. Cox, International Conference on Machine Learning, 2013, pp. 115–123.

[6] Y. Takano and K. Houk, *Journal of Chemical Theory and Computation*, 2005, **1**, 70–77.

[7] F. Neese, F. Wennmohs, U. Becker and C. Riplinger, *The Journal of Chemical Physics*, 2020, **152**,.

[8] L. Goerigk and S. Grimme, *Phys. Chem. Chem. Phys.*, 2011, **13**, 6670–6688.

[9] T. Mukhopadhyaya, T. Lee, C. Ganley, P. Clancy and H. E. Katz, *ACS Appl. Polym. Mater.*, 2022, **4**, 2065–2080.

[10] A. Mannodi-Kanakkithodi and M. K. Chan, *Energy & Environmental Science*, 2022, **15**, 1930–1949.

[11] Y. Eatmon, O. Romiluyi, C. Ganley, R. Ni, I. Pelczer, P. Clancy, B. P. Rand and J. Schwartz, *The Journal of Physical Chemistry Letters*, 0, **0**, 6130–6137.

[12] R. Mercado, R.-S. Fu, A. V. Yakutovich, L. Talirz, M. Haranczyk and B. Smit, *Chemistry of Materials*, 2018, **30**, 5069–5086.

[13] A. Deshwal, C. M. Simon and J. R. Doppa, *Molecular Systems Design & Engineering*, 2021, **6**, 1066–1086.