







**Fig. 1** Curated design space of the hypothetical DPP-based polymers studied in this work. The meaning of A, B, C is as given in the text. We used four choices of A, and nine each of B and C as denoted beneath the boxes. Note that  $n = 1$  in the simulations performed here, due to computational resource constraints.

any number of potential motifs can, in principle, be used to construct these hypothetical polymers; we chose to design our polymer using four different motifs arbitrarily. Each of the 324 possibilities we considered has different molecular electronic parameters, which we characterized by performing DFT calculations. For the DFT calculations, we elected to use the B97-D3 functional, as recommended by Goerigk *et al.* for main-group non-covalent interactions at the GGA level of theory with dispersion corrections. We combined this functional with the accurate def2-TZVP basis set that Oliveira *et al.* found to effectively balance accuracy and computational cost.<sup>37,38</sup>

A given polymer repeat unit constructed in this way is theoretically capable of being polymerized, processed, and cast into a thin-film device in conjunction with any number of potential solvents. But some solvents are known to solvate a polymer more effectively than others due to the complex relationship between the chemical properties of solute and solvent. It is this processing aspect of materials design that we wanted to explore. Due to the resources required for modeling the electronic structures of up to hundreds of atoms simultaneously, we represented each polymer as its repeat unit, *i.e.*,  $n = 1$  in Fig. 1. We thus calculated the solvation free energy change ( $\Delta G_{\text{solv}}$ ) for all 324 polymer repeat units immersed in 26 different implicitly modeled solvents using DFT. This yielded a final design space of 8424 possible polymer-solvent interactions. We acknowledge that it is a limitation of our work to have studied single-chain, infinitely-dilute, enthalpic solvation of these polymers. Moreover, due to the structural simplifications necessary for high-fidelity electronic structure modeling within a reasonable time and computational budget, our calculations do not include inter- or intra-chain interaction effects, entanglement molecular weight effects, free-volume contributions, or kinetic constraints to solvation, all of which affect the dissolution process, but at a much larger scale than is fea-

ible to model with DFT. We envision these effects could be the subject of future studies that utilize coarse-grained or classical molecular dynamics (MD) with a suitable interatomic potential. However, the focus of the present study and its general value is on the electrostatic contributions of functional groups within polymer repeat units, and which are the starting point of polymer design. The insight gained herein should therefore be used as a single tool in an arsenal of multi-scale tools for computational polymer design, which can save experimental time and cost with high-throughput screening *a priori*.

There are many different ways to represent a chemical structure in model space. Commonly used descriptor libraries include Dragon, Mordred, and RDKit.<sup>39–41</sup> However, the descriptors from these libraries are primarily topology-based: they only include the connectivity and some basic composition data. Pereira *et al.* found that including DFT-calculated orbital energy data *in addition to* connectivity descriptors improved their model's accuracy of predicting HOMO and LUMO energies.<sup>42</sup> Understanding the value of electronic structure data, we identified some features characterizing the design space that effectively encode expert chemical knowledge into the training set. A common heuristic for solvation says that “like dissolves like”. Even though this is a simplification of a very complex process, the essence of this dictum holds. The electrostatic properties (governed by the electron density) of the polymer repeat unit and solvent pair must be similar in order to optimize the free energy of solvation which, in turn, will speak to its thermodynamic stability. To this end, we elected to represent our polymers and solvents with the descriptors shown in Table 1.

These descriptors relate to the distribution of charge around the respective species and are readily calculable with the ORCA DFT software package,<sup>43,44</sup> or using a post-processing software recommended in ORCA's documentation called Multiwfn.<sup>45</sup> In Table 1, “ESP” stands for the “electrostatic potential”, and the molecular polarity index (MPI) is a normalization across the area of the autocorrelation of ESP charge over the van der Waals (vdW) surface at a constant potential value, in this case the default value  $\rho = 0.001$  a.u.<sup>46</sup> The equation for calculating MPI is:

$$\text{MPI} = \frac{1}{A} \iint_S V(\mathbf{r}) |dS| \quad (1)$$

where  $A$  is vdW surface area,  $V(\mathbf{r})$  is the value of the ESP at a point  $\mathbf{r}$  in space, and  $S$  is the molecular vdW surface.

**Table 1** Model descriptors (properties) used in this work for both polymer repeat unit (left column) and solvent (right column)

Polymer repeat unit	Solvent
Isotropic quadrupole	Isotropic quadrupole
Isotropic polarizability	Isotropic polarizability
ESP <sub>min</sub>	ESP <sub>min</sub>
ESP <sub>max</sub>	ESP <sub>max</sub>
Polarity index	Polarity index
% polar surface area	% polar surface area
HOMO–LUMO gap	Dielectric constant







greater enthalpic preference for the bound state over the isolated states.

The polymers we tested exhibit a greater “stability” in the presence of solvents with a lower dielectric constant. The average binding enthalpies, in order of increasing solvent dielectric constant, of the four explicitly modeled solvents are:  $-0.29$  eV,  $-0.29$  eV,  $-0.16$  eV, and  $-0.14$  eV. Less polar solvents are predicted to solvate the 324 polymer candidates better than more polar ones. There is a correlation with the results in Fig. 2 in that *ortho*-dichlorobenzene, with  $\epsilon \sim 9.99$ , exhibits both the highest (negative) magnitude  $\Delta G_{\text{solv}}$  and the greatest binding enthalpy. A direct comparison of this correlation is displayed in Fig. S9. This suggests that this is an “optimal” solvent range for this set of polymers.

An outcome of this validation is the similarity of results using “implicit solvent” models, like the conductor-like polarizable continuum model (CPCM) or the universal solvation model (SMD), with those of the explicitly modeled solvents. This confirms that accurate and valuable insight into solvation can be obtained using implicit solvent modeling at a fraction of the cost of explicit solvation modeling. It is possible that, with greater sampling of the polymer–explicit solvent interaction space than the four test cases explicitly modeled here, there would be an even stronger correlation. We note that the lower magnitude binding enthalpy between the highly polar solvents (water and acetonitrile) could arise because the most favorable binding sites are highly localized, whereas the polar binding sites available on the polymer are more diffuse. There are other experimental considerations at play, such as dopant miscibility, which affect solvent choice and were not considered in this study.

### 3.1 Machine-learned feature engineering

A more systematic approach to elucidate the quantitative nature of structure–property relationships in these polymers was conducted on the implicit solvation data using the “feature engineering” capability of PAL 2.0, a physics-informed Bayesian framework for materials discovery developed by the Clancy group.<sup>33</sup> Using XGBoost, a gradient-based machine learning algorithm for feature selection,<sup>49</sup> and the Least Absolute Shrinkage and Selection Operator (LASSO), a method for estimation of linear models,<sup>50</sup> we obtained a ranked list of the most important descriptors (features), as shown in Table 1, for quantifying a target material property, in this case,  $\Delta G_{\text{solv}}$ .

Feature correlation is addressed through the feature selection step, using the well-known XGBoost technique. During tree construction, correlated features compete to reduce the loss, and the algorithm typically selects one representative feature from a correlated set for early ‘splits’. Once such a feature is selected, additional correlated features tend to provide diminished marginal gain and consequently receive lower importance scores. As a result, highly correlated features are often down-weighted or excluded during importance-based ranking. This results in one of the highly correlated features being selected as important and the others being pruned from

the model. In this case, this method results in one of the correlated features being arbitrarily selected over the others, but that this does not affect interpretability at all as we know how these features are correlated and can still properly attribute physical meaning to the values selected.

Fig. 6 shows the relative importance of each model descriptor as determined by the labeled algorithms. Both LASSO and XGBoost models identify the polymer’s isotropic quadrupole moment as a descriptor with a high relative importance. This supports the observation of a different trend in  $\Delta G_{\text{solv}}$  for low-quadrupole *versus* high-quadrupole polymers exhibited in Fig. 3. Both models also identify the relative importance of the explicit polymer–solvent interaction, providing further evidence for the behavior observed in Fig. 5 and the ensuing commentary on the value of implicit *versus* explicit models of solvent interactions. A key difference between LASSO and XGBoost is that the former employs a linear model while the latter is better suited for non-linear modeling. This could explain the disparity in features identified in Fig. 6. LASSO identifies numerous descriptors with comparable importance, whereas XGBoost clearly identifies the solvent dielectric ( $\epsilon$ ) as the most important feature. It could be that LASSO simply needs more terms to be able to describe the input–output relationship; it cannot penalize too many terms in order to maintain accurate predictions. XGBoost, on the other hand, may be identifying the same non-linear dependence on  $\epsilon$  that is shown in Fig. 3. Within the vast polymer chemical space that exists, feature engineering has identified useful metrics to help uncover the complex relationship between polymer solutes and solvents.



Fig. 6 Relative descriptor importance as quantified by two different machine learning techniques, LASSO (A) and XGBoost (B).



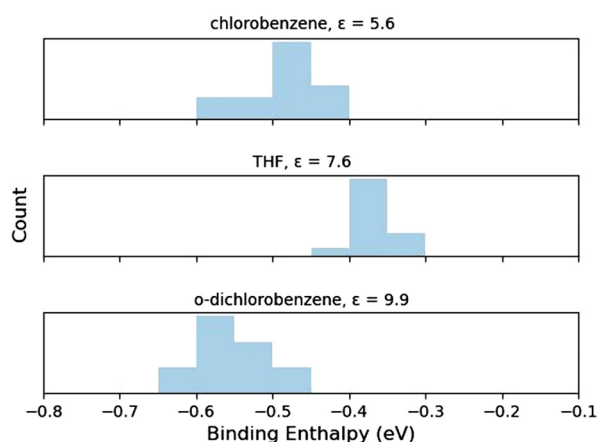




**Fig. 8** Box-plot distributions of DFT-calculated values of the Gibbs free energy of solvation for 14 IDT-based polymers in 25 implicitly modeled solvents.

solvation in chlorobenzene and *ortho*-dichlorobenzene. However, we note that the DFT-calculated  $\Delta G_{\text{solv}}$  in tetrahydrofuran (THF) is consistently higher in magnitude (less solvated) than either of its dielectric neighbors, chlorobenzene and *ortho*-dichlorobenzene. A follow-up study wherein we conduct an explicit solvation study on these three solvents, analogous to Fig. 5, was conducted for all the A-IDT combinations present in Fig. S4. These results, shown in Fig. 9, demonstrate the same trend as seen in the implicit calculations: THF has a less negative (unfavorable) binding enthalpy with these polymers than its dielectric neighbors. We posit that this may be due to the more localized dipole moment of THF as compared to chlorobenzene and *ortho*-dichlorobenzene. The larger aromatic rings present in these two dielectric neighbors also allow for greater charge stabilization, which enhances binding enthalpy.

Experimental solvation observations on an indacenodithiophene-benzothiadiazole (IDTBT) polymer revealed insolubility in a dielectric as low as acetone ( $\epsilon = 20.7$ ) while noting sol-



**Fig. 9** Distributions of explicitly modeled polymer–single solvent interaction enthalpies for 14 IDT-based polymers for THF and its dielectric neighbors. The farther to the left the average of the enthalpy distribution, the more favorable it is.

vation in THF and its neighbors. We observed a maximum  $\epsilon$  value at which a polymer is soluble; regardless of this value, solvents with a dielectric lower than this critical point exhibit solubility. An analysis of variance (ANOVA) test was conducted on the solvation data, grouped by solvent, and found a rejection of the null hypothesis: that all mean  $\Delta G_{\text{solv}}$  values among solvent groups were equal. Consequently, we used Tukey's range test to quantify the confidence intervals of the data within each group. The results of this test are shown in Fig. 10.

There may be minimal overlap in the confidence intervals of the chlorobenzene and *ortho*-dichlorobenzene  $\Delta G_{\text{solv}}$  population means but, importantly, they both exhibit a statistical difference (*i.e.*, no overlap) from THF. This conclusion is reinforced with evidence from k-means clustering analysis on the DPP data. Fig. S6 and S7 demonstrate that four clusters is optimal for converging a within-class sum of squared distances to an asymptote, when grouping solvation data by solvent, and that those four clusters occupy discrete ranges along Principal Component (PC) 1. These four clusters have been color-coded and applied as a filter to Fig. 11, whose underlying data is the same as Fig. 2.

THF is clustered separately from its dielectric neighbors, as are cyclohexanone and ethanol, hinting at the highly nonlinear behavior of  $\Delta G_{\text{solv}}$  as a function of  $\epsilon$ . It is possible that the two clusters with the highest magnitude  $\Delta G_{\text{solv}}$  (green and blue in Fig. 11) define a range of “optimal” solvents for this application. However, further investigation into the entropic effects of solvation is needed in order to assert the existence of a single “optimal” solvent and to better match experimental observations.

### 3.4 Strengths and limitations of the present approach

Ultimately, the design framework described herein possesses several key strengths which underlie its utility for materials design. The physics-informed Gaussian Process surrogate model, trained on an exhaustive computational survey of the



**Fig. 10** Results of a *post hoc* Tukey's Honestly Significant Difference test. The bars represent the 95% confidence interval of the  $\Delta G_{\text{solv}}$  of 324 DPP-based polymers in the 26 solvents, arranged by dielectric constant on the y-axis. In a pairwise comparison, a lack of overlap indicates significant difference in  $\Delta G_{\text{solv}}$ .





Using DFT-calculated properties that quantify charge density distribution as descriptors, we trained a GP model and demonstrated the usefulness of the PAL 2.0 Bayesian optimization algorithm to predict  $\Delta G_{\text{solv}}$ . The PAL 2.0 codebase found the optimum result, in its best replicate performance, within two iterations and identified, using XGBoost and LASSO methods for “feature engineering”, important descriptors needed to predict  $\Delta G_{\text{solv}}$ , our metric of success for insight into the solution behavior of the system. Explicit all-atom modeling of the polymer–solvent interactions revealed a similar trend to the behavior exhibited by  $\Delta G_{\text{solv}}$  when an implicit solvent model is used, in which the solvent is represented by a mean field of a given dielectric constant ( $\epsilon$ ).

We identified a molecular electronic parameter that demonstrates near-linear additivity when constructing an aggregate polymer from functional group sub-units, namely the isotropic quadrupole moment. The importance of this observation is that we have identified a fast route to determining better polymer candidates, without needing to resort to machine learning, as used in this PAL-generated study. It suggests that future studies of different semiconducting polymers, beyond simply DPP-based ones, could conceivably only need to model the isotropic quadrupole moment of the various functional group sub-units in the posited polymer in order to be able to predict the extent of solvation of this novel polymer. Thus, this work lays the groundwork for a unique closed-loop, hybrid computational–experimental study wherein polymers could be designed *in silico* for a target property and then synthesized and characterized with experimental techniques. Ameliorating computational property predictions with experimentally observable properties will be important for model validation and investigation design.

## Author contributions

C. G.: conceptualization, methodology, data curation, visualization, and writing. Y. F.: data curation, visualization, and writing. H. S. K.: experimental data curation and writing. H. E. K.: conceptualization, supervision, and writing. P. C.: conceptualization, funding acquisition, project administration, resources, supervision, and writing.

## Conflicts of interest

The authors declare no conflict of interest.

## Data availability

The DPP-based polymer electronic structure calculations dataset generated during this study is available on Github at <https://github.com/pclancy-lab/polymer-builder>.

Supplementary information (SI) is available. See DOI: <https://doi.org/10.1039/d5py01172h>.

Code availability: the PAL 2.0 software is available on Github at <https://github.com/ClancyLab/PAL2>.

## Acknowledgements

This work has been sponsored by an NSF (CHE) award to Drs Howard E. Katz and Paulette Clancy, award number 2107360 and by the Hopkins Institute for Data-Intensive Engineering and Science (IDIES) through the provision of a seed grant to partially fund this study. The authors gratefully acknowledge the support of both awards for this work. We thank Dr Tushita Mukhopadhyay for sharing her expertise in DPP-based polymer systems, and Dr Maitreyee Sharma Priyadarshini for assistance with the PAL 2.0 methodology. Computing resources were provided by the ARCH high-performance computing (HPC) facilities at Johns Hopkins University, which is supported by National Science Foundation (NSF) grant number OAC 1920103.

## References

- 1 A. R. Murad, A. Iraqi, S. B. Aziz, S. N. Abdullah and M. A. Brza, *Polymers*, 2020, **12**, 2627.
- 2 S. A. Lopez, B. Sanchez-Lengeling, J. De Goes Soares and A. Aspuru-Guzik, *Joule*, 2017, **1**, 857–870.
- 3 X. Yan, M. Xiong, J.-T. Li, S. Zhang, Z. Ahmad, Y. Lu, Z.-Y. Wang, Z.-F. Yao, J.-Y. Wang, X. Gu and T. Lei, *J. Am. Chem. Soc.*, 2019, **141**, 20215–20221.
- 4 J. Lee, S. A. Park, S. U. Ryu, D. Chung, T. Park and S. Y. Son, *J. Mater. Chem. A*, 2020, **8**, 21455–21473.
- 5 T.-Q. Nguyen, R. Y. Yee and B. J. Schwartz, *J. Photochem. Photobiol., A*, 2001, **144**, 21–30.
- 6 M. C. Scharber and N. S. Sariciftci, *Adv. Mater. Technol.*, 2021, **6**, 2000857.
- 7 I. Kaur, W. Jia, R. P. Kopeski, S. Selvarasah, M. R. Dokmeci, C. Pramanik, N. E. McGruer and G. P. Miller, *J. Am. Chem. Soc.*, 2008, **130**, 16274–16286.
- 8 V. Sharma, C. Wang, R. G. Lorenzini, R. Ma, Q. Zhu, D. W. Sinkovits, G. Pilania, A. R. Oganov, S. Kumar, G. A. Sotzing, S. A. Boggs and R. Ramprasad, *Nat. Commun.*, 2014, **5**, 4845.
- 9 S. Ren, Y. Ding, W. Zhang, Z. Wang, S. Wang and Z. Yi, *Polymers*, 2023, **15**, 3803.
- 10 K. Lee, M.-K. Jeong, E. H. Suh, W. Jeong, J. G. Oh, J. Jang and I. H. Jung, *Adv. Electron. Mater.*, 2022, **8**, 2101105.
- 11 NREL, *Best Research Cell Efficiency Chart*, 2024, <https://www.nrel.gov/pv/cell-efficiency.html>.
- 12 A. T. John, A. Narayanasamy, D. George and M. Hariharan, *Cryst. Growth Des.*, 2022, **22**, 1237–1243.
- 13 S. Lee, D. Jeong, C. Kim, C. Lee, H. Kang, H. Y. Woo and B. J. Kim, *ACS Nano*, 2020, **14**, 14493–14527.
- 14 D. Lungwitz, A. E. Mansour, Y. Zhang, A. Opitz, S. Barlow, S. R. Marder and N. Koch, *Chem. Mater.*, 2023, **35**, 672–681.
- 15 M. Duhandžić, M. Lu-Diaz, S. Samanta, D. Venkataraman and Z. Akšamija, *Phys. Rev. Lett.*, 2023, **131**, 248101.
- 16 W. J. R. Jayasundara and G. Schreckenbach, *J. Phys. Chem. C*, 2020, **124**, 17528–17537.



- 17 J. Tang, Y. H. Pai and Z. Liang, *ACS Energy Lett.*, 2022, **7**, 4299–4324.
- 18 F. M. A. Alzahrani, M. Sagir, M. Saqib, S. Bashir, T. Sarwar, S. Hussain, S. Murtaza, A. Mushtaq, R. Razzaq, Z. A. Alrowaili and M. S. Al-Buriahi, *Comput. Mater. Sci.*, 2024, **239**, 112961.
- 19 N. Meftahi, M. Klymenko, A. J. Christofferson, U. Bach, D. A. Winkler and S. P. Russo, *npj Comput. Mater.*, 2020, **6**, 1–8.
- 20 S. Wu, Y. Kondo, M. a. Kakimoto, B. Yang, H. Yamada, I. Kuwajima, G. Lambard, K. Hongo, Y. Xu, J. Shiomi, C. Schick, J. Morikawa and R. Yoshida, *npj Comput. Mater.*, 2019, **5**, 66.
- 21 P. Xu, T. Lu, L. Ju, L. Tian, M. Li and W. Lu, *J. Phys. Chem. B*, 2021, **125**, 601–611.
- 22 T. B. Martin and D. J. Audus, *ACS Polym. Au*, 2023, **3**, 239–258.
- 23 L. Gao, J. Lin, L. Wang and L. Du, *Acc. Mater. Res.*, 2024, **5**, 571–584.
- 24 J. Hachmann, R. Olivares-Amaya, S. Atahan-Evrenk, C. Amador-Bedolla, R. S. Sánchez-Carrera, A. Gold-Parker, L. Vogt, A. M. Brockway and A. Aspuru-Guzik, *J. Phys. Chem. Lett.*, 2011, **2**, 2241–2251.
- 25 O. H. Omar, T. Nemataram, A. Troisi and D. Padula, *Sci. Data*, 2022, **9**, 54.
- 26 T. Mukhopadhyaya, T. D. Lee, C. Ganley, S. Tanwar, P. Raj, L. Li, Y. Song, P. Clancy, I. Barman, S. Thon and H. E. Katz, *Adv. Funct. Mater.*, 2022, **32**, 2208541.
- 27 T. Mukhopadhyaya, J. Wagner, T. D. Lee, C. Ganley, S. Tanwar, P. Raj, L. Li, Y. Song, S. J. Melvin, Y. Ji, P. Clancy, I. Barman, S. Thon, R. S. Klausen and H. E. Katz, *Adv. Funct. Mater.*, 2024, **34**, 2309646.
- 28 N. E. Jackson, B. M. Savoie, K. L. Kohlstedt, M. Olvera de la Cruz, G. C. Schatz, L. X. Chen and M. A. Ratner, *J. Am. Chem. Soc.*, 2013, **135**, 10475–10483.
- 29 J. Liu, Y. Shi and Y. Yang, *Adv. Funct. Mater.*, 2001, **11**, 420–424.
- 30 S. N. Patel, A. M. Glaudell, K. A. Peterson, E. M. Thomas, K. A. O'Hara, E. Lim and M. L. Chabinye, *Sci. Adv.*, 2017, **3**, e1700434.
- 31 D. Dhabal, Z. Jiang, A. Pallath and A. J. Patel, *J. Phys. Chem. B*, 2021, **125**, 5434–5442.
- 32 B. Lukose, S. V. Bobbili and P. Clancy, *Mol. Simul.*, 2017, **43**, 743–755.
- 33 M. S. Priyadarshini, O. Romiluyi, Y. Wang, K. Miskin, C. Ganley and P. Clancy, *Mater. Horiz.*, 2024, **11**, 781–791.
- 34 M. S. Priyadarshini, E. Gienger, J. Ren, B. Piloseno, E. A. Pogue, P. K. Lambert, J. S. Ko and P. Clancy, *Machine Learning-Driven Closed-Loop Discovery of Hard Multiple Principal Element Alloys*, 2025, <https://chemrxiv.org/engage/chemrxiv/article-details/68a7f1d1a94eede154053d56>.
- 35 X. Zhang, H. Bronstein, A. J. Kronemeijer, J. Smith, Y. Kim, R. J. Kline, L. J. Richter, T. D. Anthopoulos, H. Sirringhaus, K. Song, M. Heeney, W. Zhang, I. McCulloch and D. M. DeLongchamp, *Nat. Commun.*, 2013, **4**, 2238.
- 36 R. Chen, T. Jin, Y. Liu, T. Zhang, X. Liu, L. Zhang, Y. Chen, H. Li, X. Duan and Y. Han, *Macromolecules*, 2023, **56**, 5356–5368.
- 37 L. Goerigk and S. Grimme, *Phys. Chem. Chem. Phys.*, 2011, **13**, 6670–6688.
- 38 M. T. d. Oliveira, J. M. A. Alves, N. L. Vrech, A. A. C. Braga and C. A. Barboza, *Phys. Chem. Chem. Phys.*, 2023, **25**, 1903–1922.
- 39 I. V. Tetko, J. Gasteiger, R. Todeschini, A. Mauri, D. Livingstone, P. Ertl, V. A. Palyulin, E. V. Radchenko, N. S. Zefirov, A. S. Makarenko, V. Y. Tanchuk and V. V. Prokopenko, *J. Comput.-Aided Mol. Des.*, 2005, **19**, 453–463.
- 40 H. Moriwaki, Y.-S. Tian, N. Kawashita and T. Takagi, *J. Cheminf.*, 2018, **10**, 4.
- 41 G. Landrum, P. Tosco, B. Kelley, R. Rodriguez, D. Cosgrove, R. Vianello, P. sriniker, G. Jones, N. Schneider, E. Kawashima, D. Nealschneider, A. Dalke, M. Swain, B. Cole, S. Turk, A. Savelev, A. Vaucher, M. Wójcikowski, I. Take, tadhurst cdd, V. F. Scalfani, R. Walker, D. Probst, K. Ujihara, J. Lehtivarjo, G. Godin, A. Pahl, F. Bérenger and H. Faara, *rdkit/rdkit:2024\_09\_3 (Q3 2024) Release*, 2024, <https://zenodo.org/records/14245989>.
- 42 F. Pereira, K. Xiao, D. A. R. S. Latino, C. Wu, Q. Zhang and J. Aires-de-Sousa, *J. Chem. Inf. Model.*, 2017, **57**, 11–21.
- 43 F. Neese, *Wiley Interdiscip. Rev.:Comput. Mol. Sci.*, 2011, **2**, 73–78.
- 44 F. Neese, *Wiley Interdiscip. Rev.:Comput. Mol. Sci.*, 2017, **8**, e1327.
- 45 T. Lu and F. Chen, *J. Comput. Chem.*, 2012, **33**, 580–592.
- 46 Z. Liu, T. Lu and Q. Chen, *Carbon*, 2021, **171**, 514–523.
- 47 A. V. Marenich, C. J. Cramer and D. G. Truhlar, *J. Phys. Chem. B*, 2009, **113**, 6378–6396.
- 48 V. Barone and M. Cossi, *J. Phys. Chem. A*, 1998, **102**, 1995–2001.
- 49 T. Chen and C. Guestrin, Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, New York, NY, USA, 2016, pp. 785–794.
- 50 R. Tibshirani, *J. R. Stat. Soc. Ser. B: Methodol.*, 1996, **58**, 267–288.
- 51 F. Archetti and A. Candelieri, *Bayesian Optimization and Data Science*, Springer International Publishing, Cham, 2019, pp. 57–72.
- 52 Y. Zheng, S. Zhang, J. B.-H. Tok and Z. Bao, *J. Am. Chem. Soc.*, 2022, **144**, 4699–4715.
- 53 J. Imbrogno, K. Maruyama, F. Rivers, J. R. Baltzegar, Z. Zhang, P. W. Meyer, V. Ganesan, S. Aoshima and N. A. Lynd, *ACS Macro Lett.*, 2021, **10**, 1002–1007.
- 54 S. Raguraman, M. S. Priyadarshini, T. Nguyen, R. McGovern, A. Kim, A. J. Griebel, P. Clancy and T. P. Weihs, *J. Magnesium Alloys*, 2024, **12**, 2267–2283.

