A Physics-Infused Deep Learning Model for the Prediction of Refractive Indices and Its Use for the Large-Scale Screening of Organic Compound Space

Mojtaba Haghighatlari,[1,*] Gaurav Vishwakarma,[1] Mohammad Atif Faiz Afzal,[1] and Johannes Hachmann[1,2,3,†]

[1] *Department of Chemical and Biological Engineering, University at Buffalo, The State University of New York, Buffalo, NY 14260, United States*

[2] *Computational and Data-Enabled Science and Engineering Graduate Program, University at Buffalo, The State University of New York, Buffalo, NY 14260, United States*

[3] *New York State Center of Excellence in Materials Informatics, Buffalo, NY 14203, United States*

We present a multitask, physics-infused deep learning model to accurately and efficiently predict refractive indices (RIs) of organic molecules, and we apply it to a library of 1.5 million compounds. We show that it outperforms earlier machine learning models by a significant margin, and that incorporating known physics into data-derived models provides valuable guardrails. Using a transfer learning approach, we augment the model to reproduce results consistent with higher-level computational chemistry training data, but with a considerably reduced number of corresponding calculations. Prediction errors of machine learning models are typically smallest for commonly observed target property values, consistent with the distribution of the training data. However, since our goal is to identify candidates with unusually large RI values, we propose a strategy to boost the performance of our model in the remoter areas of the RI distribution: We bias the model with respect to the under-represented classes of molecules that have values in the high-RI regime. By adopting a metric popular in web search engines, we evaluate our effectiveness in ranking top candidates. We confirm that the models developed in this study can reliably predict the RIs of the top 1,000 compounds, and are thus able to capture their ranking. We believe that this is the first study to develop a data-derived model that ensures the reliability of RI predictions by model augmentation in the extrapolation region on such a large scale. These results underscore the tremendous potential of machine learning in facilitating molecular (hyper)screening approaches on a massive scale and in accelerating the discovery of new compounds and materials, such as organic molecules with high-RI for applications in opto-electronics.

I. INTRODUCTION

In recent years, data-driven approaches leading to big data scenarios are rapidly gaining momentum in the chemical and materials domain [1]. A wealth of information can be extracted from such data using a variety of machine learning (ML) techniques. ML prediction models are increasingly being used as surrogates for physics-based models and can reveal intricate and often hidden structure-property relationships. These models are strikingly faster than their physics-based counterparts with little compromise in terms of accuracy [2, 3]. Thus, ML models can be applied as a catalyst in the process of virtual high-throughput screening to expedite the exploration of molecular space (see, e.g., Refs. [4–11]). Currently, this strategy is presumed as the backbone for the accelerated discovery of materials with tailored properties.

The optical properties of materials are of key importance for a range of optic and optoelectronic applications, such as organic light-emitting diodes, photovoltaics, image sensors [12, 13]. A high index of refraction (RI) is often one desirable property, in particular for lens components. In a previous study, we explored the utility of organic compounds in this regard and presented a computational protocol to model their RI [14–17]. This protocol was cast to the virtual high-throughput screening of thousands of organic molecules to collect data of associated properties including static polarizability, number density (or density) and RI values. We employ this data set to develop ML models that efficiently and accurately predict properties of unlabeled molecules.

There is a rich history of developing ML models to facilitate the prediction of RI values. These studies have been applied to a wide range of materials from organic aerosols and common organic compounds to different classes of polymers [18–21]. However, a majority of these models are either trained on a limited data set, or only apply linear models, which often lack enough complexity to capture the underlying characteristics of data. Thus, we combine the recent developments in the data mining area with the large data set from our computational studies to advance the modeling in terms of the accuracy and reliability for the prediction of a significantly larger molecular library. In addition, the top candidates resulting from our study are validated by performing the previously established computational protocol that was used for generating the data [14].

In this paper, we first detail the generation of reference data (Sec. II). The data generation process includes molecular library generation, and application of virtual high-throughput screening to compile corresponding properties of the resulting library. Next, we describe methods employed in our data-derived modeling, includ-

* mojtabah@buffalo.edu

† hachmann@buffalo.edu

ing feature representation (Sec. II B), standard and customized neural network architectures and their training details (Sec. II C), and transfer learning approach to outsmart the ML models (Sec. II D). We also propose a fine-tuning strategy to assure the reliability of predictions for high-RI candidates (Sec. II E). Sec. III presents outcomes of the modeling, and emphasizes the improved extrapolation approach. We discuss most important observations or otherwise justify exceptions in Sec. IV. Our findings are summarized in Sec. V.

II. BACKGROUND, METHODS, AND COMPUTATIONAL DETAILS

A. Study of High-Refractive-Index Candidates

As described in a recent study [22], we generated a molecular library of 1.5 million small organic molecules, using our combinatorial library generator package, *ChemLG* [23, 24]. The library is constructed from 15 molecular building blocks and is constrained by molecular weight and number of ring-moieties per molecule. In addition, we employ the Lorentz-Lorenz equation to calculate the RI (n_{r}) of the molecules as a function of their polarizability (α) and number density (N). The Lorentz-Lorenz equation is given by

$$n_{\mathrm{r}} = \sqrt{\frac{1 + 2\alpha N/3\varepsilon_0}{1 - \alpha N/3\epsilon_0}}.$$

We recently developed an accurate modeling protocol [16, 22, 25] to compute α and N using the Kohn-Sham density functional theory (DFT) [26, 27] and molecular dynamics (MD) simulation, respectively. Since these computational studies of the entire library of 1.5 million molecules is time- and resource-intensive, we limit the screening to a random subset of 100,000 molecules. The resulting data set serves as the ground-truth for our data modeling to accelerate screening of the entire library by predicting corresponding properties, i.e., n_{r}, α, and N.

The DFT calculations for 100,000 molecules are carried out using the PBE0 functional [28] and double-ζ def2-SVP basis set [29] along with D3 dispersion correction. Due to the dependency of derived RI predictions to the employed quantum computations for polarizability values, we also perform calculations using the larger basis set, triple-ζ def2-TZVP [29]. However, considering the computational cost of this extra calculations at higher quantum chemistry approximations, we only use the triple-ζ basis set for additional calculations of a random subset of 10,000 molecules (out of 100,000). We use this data set to augment our data-derived models without performing an exhaustive calculations for 100,000 molecules (see Sec. II D). Note that the new polarizability velues are used for the calculation of RI values using LL equation with the same density values from MD simulation. In this paper, the def2-SVP and def2-TZVP basis sets and corresponding data sets are abbreviated as SVP and TZVP for convenience. The statistics for both data sets are also provided in the Supplementary Material (Tables S1 and S2).

B. Feature Representation

We use hand-crafted molecular descriptors to provide numerical representation of molecules [30, 31]. Two families of descriptors that we use in this study are: (1) topological and physicochemical features from Dragon 7 [32], and (2) molecular fingerprints (FP). All these descriptors are based on a molecular graph, i.e., a 2-dimensional representation that depends only on atom types and connectivities. A SMILES representation [33] of molecules provides adequate chemical information to calculate these types of descriptors. The total number of Dragon descriptors after pre-processing (e.g., removing constant columns) reduces to 1893 features. In order to represent molecules using FP, we encode molecules to binary vectors using three FP algorithms contained within RDKit [34]. These are Morgan FP with circular radius 2 [35, 36], hashed topological torsion FP (HTT) [37], and hashed atom pair FP (HAP) [38]. The length of all FP vectors is set to 1024 bits.

C. Regression Models

In this study, we focus on deep neural networks (DNN) for the regression task [39]. The choice of DNNs for modeling is due to their promising performance and flexibility in the design of neural network architectures [40]. We use standard fully-connected DNN (Fig. 1a) to train one model for each of three properties using one of descriptor sets (i.e., resulting in 12 single DNN models in total). To account for the effects of hyperparameters that dictate the traininig and architecture of models, e.g., number of hidden layers, number of neurons, learning rate, regularization parameter, etc., a fully customizable genetic algorithm (GA) [41] code is developed whereby we efficiently navigate through the hyperparameter space to optimize the training models. We perform this ML work using *ChemML* [23, 42, 43], our program package for machine learning and informatics in chemical and materials research.

The data set is divided randomly into training and test sets with a 9:1 ratio. We use the training data to fit models and tune hyperparameters based on their evaluation on a 10% hold-out validation set. In addition to unbiased data set splits, we apply early stopping, dropout, and l2 regularization parameter to avoid overfitting [44, 45]. The best set of hyperparameters is used to define the final models. To assess the dependency of these models to the size of training set, we train each model on incrementally increasing ratios from 5% to 100% and evaluate

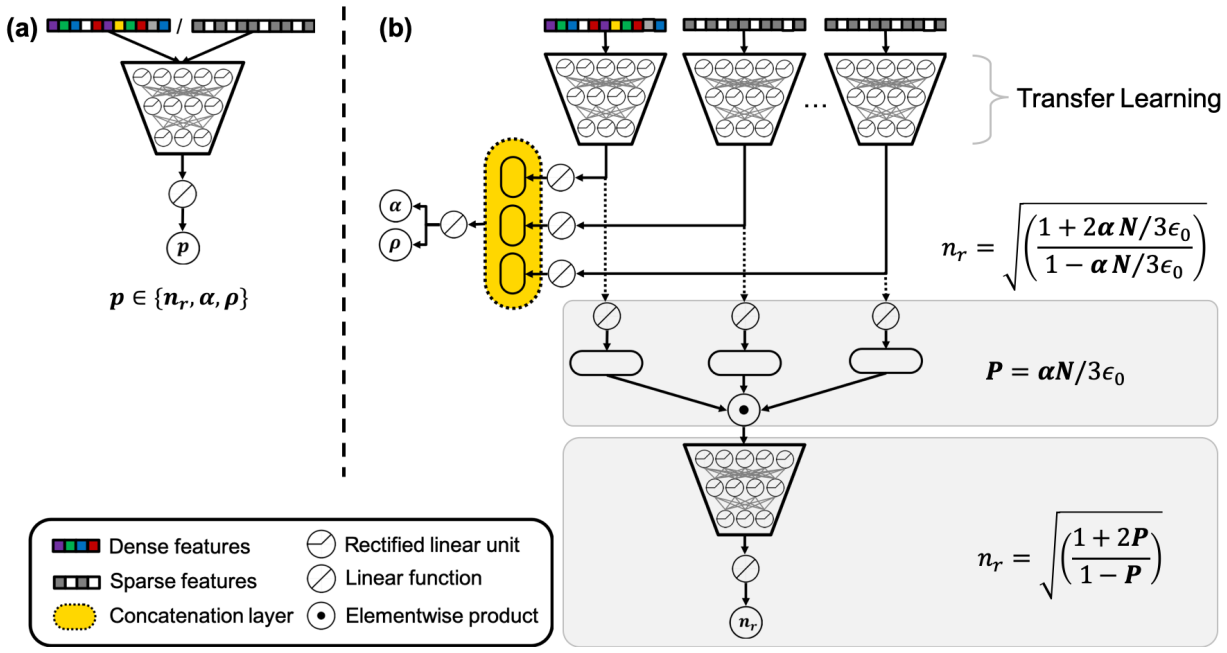

FIG. 1: Schematic of the models to predicting density (ρ), polarizability (α), and refractive index (n_r) of organic molecules. We first develop (a) standard deep neural networks (DNN)) for each property and representation (12 models in total). Based upon the insights from standard models, (b) we design a multitask physics-infused DNN to utilize the combination of dense (real values), and sparse (binary bits) feature representations. The top hidden layers of the multitask model can also be used in the transfer learning approach to leverage training on small size data sets.

them on the hold-out test set. We compare predictions of these models with calculated properties using following metrics: mean absolute error (MAE), root mean squared error (RMSE), mean absolute percentage error (MAPE), and regression correlation coefficient (R^2).

Based on the results of standard DNNs trained on each of the descriptors, we design a new model that improves the prediction performance by: (1) multitask learning [46], i.e., learning all three properties using one model, (2) utilizing all the calculated feature sets, and (3) imitating the Lorentz-Lorenz equation in the architecture of the DNNs. The overall model structure (Fig. 1b) is designed to first independently transform all four input feature sets to the latent space provided by multilayer DNN structure. This step facilitates merging the dense Dragon features with sparse FP features [47]. Next, we concatenate and linearly map the corresponding latent space to the α and N. This way we make sure that the transformed features, as latent space, are tuned to reproduce the two essential properties for calculation of n_r. The key to the design of this model is that each descriptor shows a different performance for the prediction of available properties. Therefore, for the final prediction of the n_r, we first multiply the elements of the latent layers by each other to imitate the product of α and N in the Lorentz-Lorenz equation. We finalize the prediction with one last fully-connected network to learn the n_r.

D. Transfer Learning

It is generally accepted that additional training data often improves the performance of ML predictive models, unless they are saturated in which case they exhibit a constant performance. Therefore, one would expect that predictive models trained on 100,000 training data at double-ζ-quality (SVP basis set) have a higher accuracy as compared to those trained on the smaller but more accurate 10,000 data points at the triple-ζ-quality (TZVP basis set). In order to replicate the accuracy observed in quantum chemistry in ML models, one solution is based on the idea of transfer learning (TL) for DNN models [48]. TL suggests that tuned parameters from a high-quality ML model (e.g., trained on a large data set) can be reused in the structure of a new ML model to learn the essence of the high-quality but small data sets [49]. Thus, we use the TL approach to obtain desired prediction accuracy on the TZVP data.

For the purpose of TL, we transfer tuned parameters from the initial layers of the best predictive model for the SVP data set (see Fig. 1.b) to a same model with randomly initialized parameters. These parameters will be set and frozen at the equivalent layer of the new DNN structure for the training of the TZVP data set. We further optimize number of transformed hidden layers as one of the hyperparameters in the model selection task.

E. Extrapolation to 1.5 Million Molecules

In addition to the development of accurate ML models to predict RI values, we further study if the overall prediction error is applicable to the entire range of RI values in the training set. For instance, we analyze the MAE of the predictions at the tail of the RI histogram, i.e., the desired remoter areas in the RI distribution. If the prediction error in those regions of the molecular candidates is worse than the average, we fine-tune predictive models so that they are able to capture the essence of those desired underrepresented class of molecules [50]. To perform fine-tuning (FT), we carefully retrain the best model on the data points that are close to the tail of the RI distribution, i.e., those that most probably deviate from their predicted values more than the overall MAE. Note that for this study, a significantly smaller learning rate is required to avoid causing disturbance to the model.

In order to assess our FT strategy, and to find out which modeling approach provides the best ranking of the top candidates, we employ discounted cumulative gain (DCG) ranking metric [51, 52]. The DCG has its roots in the ranking of results from web search engines. It is a measure of the ranking quality for the ordered elements of the list based on a reference relevance value. We compute the DCG as

$$DCG_p = \sum_{i=1}^{p} \frac{rel_i}{\log_2(i+1)},$$

where p is the rank position of the molecules based on their RI predictions, and rel is the relevance scale of the molecules. We select top 1,000 molecules (out of 1.5 million) based on the predictions of all three modeling approaches, i.e., multitask model trained on SVP data, TL model trained on TZVP data, and FT model trained on SVP data. The derived RI values in the SVP and TZVP data sets serve as a scoring measure to compute the rel. We use the reverse ranking of the top 1,000 molecules based on their reference RI values from data sets as our graded relevance scores. Using the DCG measure, we compare the ranking quality of predictions against the ground-truth LL calculations for this study.

III. RESULTS

The learning curves of the standard DNN models (Fig. 1 a) for each descriptor set, and the multitask Lorentz-Lorenz-equiation-infused model (Fig. 1b) are displayed in the Fig. 2 for three properties of the SVP data set. The learning curves show the MAE as a function of the training set size evenly spaced from 5% to 100% ratio. The best results at 100% ratio are shown in Table I. In addition, We report the performance of best models in terms of other metrics in Supplementary Material (Tables S8-S10). We note that the MAE for the Dragon model is the best among the standard DNN models with respect to all three properties. However, the models that are derived from the FP descriptor sets outperform at least one of the other FP models for the prediction of one of the properties. In other words, Morgan and HTT models are the best FP model for the prediction of RI and polarizability, respectively. The HAP model also outperforms Morgan for the prediction of densities.

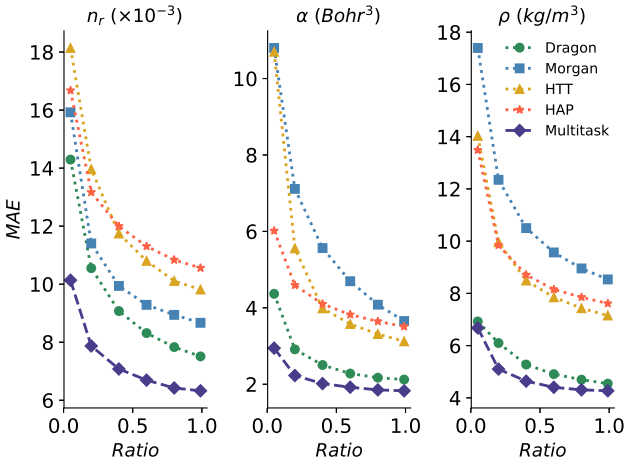

FIG. 2: The learning curves show the dependency of the mean absolute error (MAE) to the training set ratio for each model. The legend corresponds to the descriptor used for the training of the single-output standard deep neural network (DNN) models, and the new multitask physics-infused DNN model, which is trained on all four descriptor sets. All models are trained and validated on the training set from 100,000 data points (SVP data set).

The new multitask model offers a substantial improvement in terms of prediction accuracy for all three properties. In particular, the lowest MAE for the RI predictions becomes 0.006 with the correlation coefficient, $r^2 = 0.99$. This is 20% improvement compared to the best standard DNN model. Most importantly, we find that the DNN prediction errors for RI values are significantly smaller than the calculation errors (by a factor of 2 to 3) reported in our previous benchmark studies [14].

Fig. 3 shows the learning curves for the TL approach. The curves for the SVP model is same as the multitask model trained on the SVP data set. Note that here we increase the number of training set sizes at smaller proportion of the training data to match the smaller size of the training data available *via* TZVP data set. If we do not utilize the TL and train a multitask network on the 10,000 TZVP data set, the prediction errors are very similar to the SVP model. However, due to the small size of data the performance is relatively poor compared to the best model trained on the SVP data set. When TL

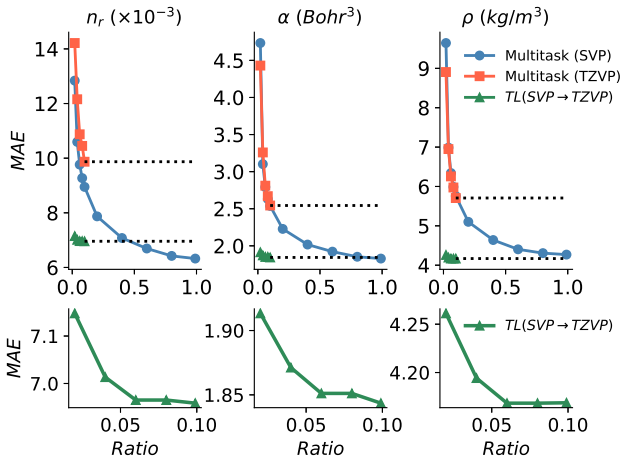

FIG. 3: The improvement in terms of reduced mean absolute error (MAE) as a function of training set ratio, by applying transfer learning (TL) approach to the multitask models. The SVP refers to the 100,000 training data with double-ζ quality. The TZVP refers to the 10,000 training data with triple-ζ quality. The green learning curves on all plots (zoomed on bottom row) show the performance of the multitask model trained with TL approach.

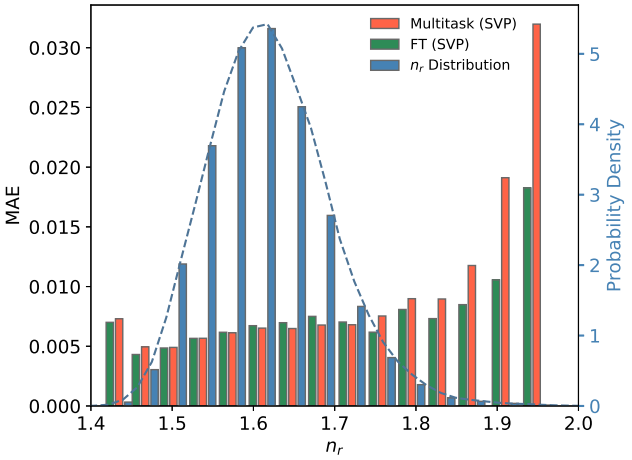

FIG. 4: The range and distribution (right y-axis) of the RI values in the hold-out test set. The mean absolute error (MAE) of predictions is shown for the evenly spaced segments of the histogram. Using the fine-tuning (FT) approach, we are able to improve the prediction error for the underrepresented molecules close to tails of the distribution. Both models are trained on the 100,000 data at double-ζ-quality (SVP).

is applied, the MAE for the same training set ratios are significantly decreased. The transferred models are able to reproduce the TZVP results by approximately 45% improvement for all properties.

The histogram and distribution of the calculated RI values in the hold-out test set of the SVP data is shown in Fig. 4. The majority of molecules have RI values between 1.5 and 1.7 with an average of 1.62. If we segment and sort the absolute prediction errors of SVP model, the resulting histogram shows exponential increase in the uncertainty of the model at the remoter areas. The MAE for the farthest segments associated with highest RI values, is approximately 5 times bigger than the overall MAE of the model. After fine-tuning, top candidates can be approximated two times more accurately than the original model. We also note that the prediction errors in the distribution peak is slightly increasing, but it is insignificant to cause any change to the ranking of molecules with RI greater than 1.8.

The results of the DCG metric is shown in Fig. 5. Higher values demonstrate a better compatibility between the rankings by reference calculations and those by the predictive models. The DCG for fine-tuned model has the greatest value among all of the models. The ranking of top candidates based on the referenced RI values in SVP data is closest to the FT SVP model, original multitask SVP model, and TL TZVP model, sequentially. If we consider the TZVP data as reference, the FT SVP model still aligns better compared to the other two models. However, the ranking by TL TZVP model is slightly better correlated with the reference ranking, which is expected due to the same level of data quality.

TABLE I: Overall prediction error of models in terms of mean absolute error (MAE). The table summarizes Fig. 2 at 100% training set ratio.

model	n_r	α ($Bohr^3$)	ρ (kg/m^3)
Morgan	0.0087	3.65	8.53
HTT	0.0098	3.12	7.15
HAP	0.0105	3.51	7.61
Dragon	0.0076	2.12	4.54
Multitask	0.0063	1.83	4.27

IV. DISCUSSION

The benefit of utilizing 2D descriptors in this study is the computational efficiency of their extraction. Thereby, we can avoid the time-intensive geometry optimization for the molecules. Moreover, we take advantage of flexibility in the design of standard DNNs to infuse the available prior knowledge into the model. The architecture design enables the interpretability, in addition to improvement in the overall performance of the model. However, the down side of working with DNNs is the exorbitant number of hyperparameters, i.e., non-trainable parameters that are required to define a model. We tackle this challenge by using our in-house evolutionary algorithms for the model selection task. We note that the

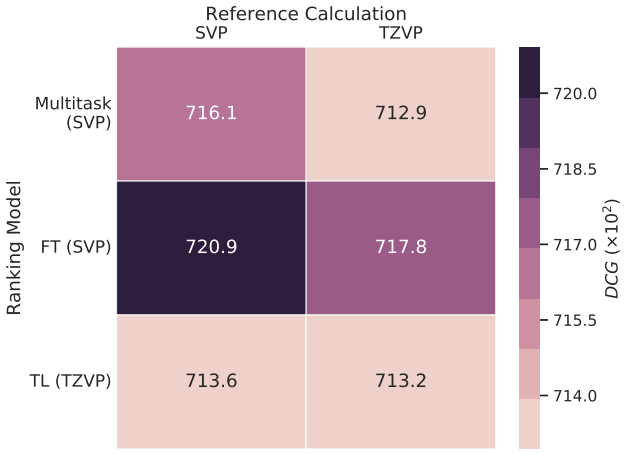

FIG. 5: The discounted cumulative gain (DCG) measures the ranking quality of the high-refractive-index candidates (i.e., top 1000 molecules) based on the predictions of data-derived models. In this regard, the rank of molecules based on original multitask model, fine-tuned model (FT), and transferred model (TL) are compared against the ranking based on polarizability calculations at double-ζ basis set (SVP) and triple-ζ basis set (TZVP) quality.

process of hyperparameter optimization plays a nontrivial role in the final model's performance. We report the best set of hyperparameters in the Supplementary Material (Table S3-S7).

The performance of the single-output standard DNNs show that each of the descriptors helps to capture structure property relationships to a certain degree, which highly depends on the target property as well. This observation motivates the idea to take advantage of a combination of features for the development of ML models. However, two challenges involved with this task are: (1) curse of dimensionality, and (2) different levels of sparsity between feature sets. We overcome both of these challenges by utilizing the deep structure of neural networks as the feature transformation step. Taking into account the availability of the essential properties for the RI calculations, the rest of the network is processing the transformed features to simultaneously learn all the properties similar to the RI calculation *via* the Lorentz-Lorenz equation. We also conduct principal component analysis (PCA) on the final hidden layers of each model before they are mapped to the target properties (see Supplementary Material Fig. S1). The differentiability of the target properties by PCA components clearly explains the role of different stages of the multitask model in learning corresponding properties. Note that using our GA code we also optimize the architecture of the models. Thereby, all the models can have same number of trainable parameters. Eventually, the multitask model is not a significantly bigger model compared to the standard DNN models. Therefore, the 20% improvement in the prediction accuracy of the multitask model is explainable based on these new advancements that we consider in its design.

Similar data mining efforts in literature have recorded an R^2 of 0.91 with a multi-linear regression model for 126 organic compounds with 5 descriptors [21], and a mean absolute error of 0.01 for another multi-linear regression model for 111 secondary organic aerosols [18]. Trained on a much larger data set, our best model, on the other hand, has a significantly better accuracy than that achieved by previously reported ML approaches.

The design of the multitask model also allows us to utilize hidden layers of tuned models (i.e., the latent space) for application in TL. Based on the learning curves of the transferred models, we understand that even half of the size of the TZVP data set is sufficient to achieve the best performance of the model. Using the TL approach, we can potentially save 90% of the calculations at triple-ζ-quality to match the prediction errors of the trained model on larger SVP data set. This amount for RI is approximately 45%. The difference can be due to error propagation in the Lorentz-Lorenz model. Accounting for the computational cost of the higher quality DFT models, this amount of reduction in the calculations is a significant improvement, specifically for the virtual-high-throughput screening that requires cost-effective approaches. Note that the density values are independent of the DFT calculations, and the little improvement in their performance is in the order of the standard deviation of their predictions.

By training on thousands of molecules covering a broad range of combinatorial structures, the multitask model is expected to make informed predictions of the entire range of target properties. However, we observe that common RI values are easier to predict due to the pervasiveness of building blocks (or specific pairs of building blocks) in the top 10% candidates. We previously studied [16] the Z-scores of the building block combinations and the results clearly prove the over/under-representation of particular substructures in both tails of the RI distribution.

Based on the DCG scores, we observe that the ranking quality of the top 1,000 candidates is best along with the fine-tuned model refer to both of the reference calculations by two different bases sets, i.e., def2-SVP and def2-TZVP. Thus, we confirm that the fine-tuning method is a very effective method to assess the properties of high-RI candidate molecules. We also note that the TZVP (TL) model provides more relevant predictions than the original SVP model to the TZVP reference, and vice verse. Therefore, it is important that what level of theory is considered to best estimate the real-world properties (i.e., ideal ranking) and thus it should be used for the training of the data-derived models.

We should also mention that the DCG scores for all the models are very close to each other, i.e., their ratio is close to one. This means that these models are essentially at the same level of prediction accuracy. However, further optimization of the original model (e.g., *via* fine-

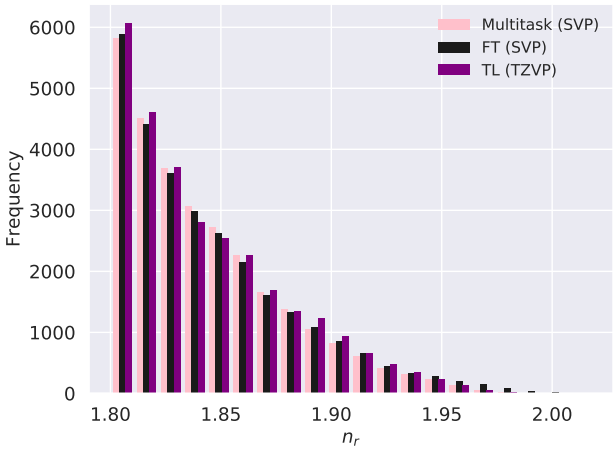

FIG. 6: The number of molecules in the 1.5 million molecular library that are among top candidates with $n_r > 1.8$. The refractive indices (n_r) are approximated using developed predictive models in this study.

tuning or TL) helps to accurately shortlist the top candidates to a limited number of molecules (e.g., less than five molecules). This way we can provide top candidates for the expensive and time-consuming experimental synthesis. Lastly, we show the number of molecules with an ambitious filter of n_r greater than 1.8 in the entire 1.5 million molecular library (Fig. 6). The RI values are predicted using the three ML models developed in this study. The total number of molecules counted in this figure is close to 30,000, which covers only 2% of the entire screening library of 1.5 million small organic molecules. The SMILES representation of the top 100 candidates based on the TZVP calculations with their corresponding predictions are available in the Supplementary Material.

V. CONCLUSIONS

In the work presented here, we designed a DNN model that incorporates the underlying physics of the given problem (i.e., the prediction of RI values *via* the Lorentz-Lorenz equation) in its architecture and merges different descriptor spaces that each have distinct benefits. We demonstrated that our model is able to reproduce all target properties (i.e., n_r, α, ρ) of the molecules in the given data set with high accuracy, and significantly outperforms the existing state-of-the-art approaches for similar problems. Next, we utilized a TL approach to further increase the model's accuracy using only a relatively small amount of high-quality data (i.e., associated with only modest additional cost). TL allows us to obtain high-level models with only a fraction of the high-level data needed for direct ML, which can thus dramatically reduce the bottleneck of the associated training data generation. Although TL has caused excitement in the ML community, it has (to the best of our knowledge) never been employed to improve the ranking of unlabeled molecules as shown in this study. We found that the transferred model is slightly better than the original in capturing the order of top-candidates. In addition to pursuing excellent overall performance, we also addressed the reliability of the model for the prediction of high-RI values (i.e., properties of top candidates at the edges of the model's applicability domain). The proposed fine-tuning approach recognizes the under-representation of training data in the extreme value range, allows the model to learn from top candidates, and thus balances the oversampling of the majority compound classes. Our fine-tuning strategy employs DCG scoring to rank molecules at all levels of theory. We conclude that the fine-tuning of ML models with data from extreme value regions is necessary to ensure a successful screening of molecular space for compounds with exceptional properties.

SUPPLEMENTARY MATERIAL

Electronic supplementary material accompanies this paper and is available through the journal website. It provides statistical analysis of all data sets that are used in this study (Table S1-S2), and tuned hyperparameter values for trained models (Table S3-S7). We also give detailed definitions of all statistical metrics used in this work along with their associated values (Table S8-S10). The principal component analysis of the final hidden layers of the trained models are also illustrated in Fig. S1.

COMPETING FINANCIAL INTERESTS

The authors declare to have no competing financial interests.

ACKNOWLEDGMENTS

This work was supported by the National Science Foundation (NSF) CAREER program (grant No. OAC-1751161), and the New York State Center of Excellence in Materials Informatics (grants No. CMI-1140384 and CMI-1148092). Computing time on the high-performance computing clusters '*Rush*', '*Alpha*', '*Beta*', and '*Gamma*' was provided by the UB Center for Computational Research (CCR). The work presented in this paper is part of MH's PhD thesis [53]. MH gratefully acknowledges support by Phase-I and Phase-II Software Fellowships (grant No. ACI-1547580-479590) of the National Science Foundation (NSF) Molecular Sciences Software Institute (grant No. ACI-1547580) at Virginia Tech [54, 55]. We are grateful to Prof. Chong Cheng for valuable discussions and insights.

REFERENCES

[1] Hachmann, J.; Windus, T. L.; McLean, J. A.; Allwardt, V.; Schrimpe-Rutledge, A. C.; Afzal, M. A. F.; Haghighatlari, M. *Framing the role of big data and modern data science in chemistry*; 2018.

[2] Butler, K. T.; Davies, D. W.; Cartwright, H.; Isayev, O.; Walsh, A. Machine learning for molecular and materials science. *Nature* **2018**, *559*, 547–555.

[3] Haghighatlari, M.; Hachmann, J. Advances of machine learning in molecular modeling and simulation. *Current Opinion in Chemical Engineering* **2019**, *23*, 51–57.

[4] Hautier, G. Finding the needle in the haystack: Materials discovery and design through computational ab initio high-throughput screening. *Computational Materials Science* **2019**, *163*, 108 – 116.

[5] Sokolov, A. N.; Atahan-Evrenk, S.; Mondal, R.; Akkerman, H. B.; Sánchez-Carrera, R. S.; Granados-Focil, S.; Schrier, J.; Mannsfeld, S. C. B.; Zoombelt, A. P.; Bao, Z.; Aspuru-Guzik, A. From computational discovery to experimental characterization of a high hole mobility organic crystal. *Nature Communications* **2011**, *2*, 437–438.

[6] 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*, 2241–2251.

[7] Olivares-Amaya, R.; Amador-Bedolla, C.; Hachmann, J.; Atahan-Evrenk, S.; Sánchez-Carrera, R. S.; Vogt, L.; Aspuru-Guzik, A. Accelerated computational discovery of high-performance materials for organic photovoltaics by means of cheminformatics. *Energy & Environmental Science* **2011**, *4*, 4849–4861.

[8] Amador-Bedolla, C.; Olivares-Amaya, R.; Hachmann, J.; Aspuru-Guzik, A. In *Informatics for materials science and engineering: Data-driven discovery for accelerated experimentation and application*; Rajan, K., Ed.; Amsterdam: Butterworth-Heinemann, 2013; Chapter 17, pp 423–442.

[9] Hachmann, J.; Olivares-Amaya, R.; Jinich, A.; Appleton, A. L.; Blood-Forsythe, M. A.; Seress, L. R.; Roman-Salgado, C.; Trepte, K.; Atahan-Evrenk, S.; Er, S.; Shrestha, S.; Mondal, R.; Sokolov, A.; Bao, Z.; Aspuru-Guzik, A. Lead candidates for high-performance organic photovoltaics from high-throughput quantum chemistry - the Harvard Clean Energy Project. *Energy & Environmental Science* **2014**, *7*, 698–704.

[10] Pyzer-Knapp, E. O.; Suh, C.; Gómez-Bombarelli, R.; Aguilera-Iparraguirre, J.; Aspuru-Guzik, A. What is high-throughput virtual screening? A perspective from organic materials discovery. *Annual Reviews of Materials Research* **2015**, *45*, 195–216.

[11] 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. *Scientific Data* **2016**, *3*, 160086.

[12] Higashihara, T.; Ueda, M. Recent progress in high refractive index polymers. *Macromolecules* **2015**, *48*, 1915–1929.

[13] Macdonald, E. K.; Shaver, M. P. Intrinsic high refractive index polymers. *Polymer International* **2014**, n/a–n/a.

[14] Afzal, M. A. F.; Cheng, C.; Hachmann, J. Combining first-principles and data modeling for the accurate prediction of the refractive index of organic polymers. *The Journal of Chemical Physics* **2018**, *148*, 241712.

[15] Afzal, M. A. F.; Hachmann, J. Benchmarking DFT approaches for the calculation of polarizability inputs for refractive index predictions in organic polymers. *Physical Chemistry Chemical Physics* **2019**, *21*, 4452–4460.

[16] Afzal, M. A. F.; Haghighatlari, M.; Prasad Ganesh, S.; Cheng, C.; Hachmann, J. Accelerated Discovery of High-Refractive-Index Polyimides via First-Principles Molecular Modeling, Virtual High-Throughput Screening, and Data Mining . *The Journal of Physical Chemistry C* **2019**, *123*, 14610–14618.

[17] Afzal, M. A. F. From virtual high-throughput screening and machine learning to the discovery and rational design of polymers for optical applications. Ph.D. thesis, University at Buffalo, 2018.

[18] Redmond, H.; Thompson, J. E. Evaluation of a quantitative structure–property relationship (QSPR) for predicting mid-visible refractive index of secondary organic aerosol (SOA). *Physical Chemistry Chemical Physics* **2011**, *13*, 6872–6882.

[19] Yu, X.; Yi, B.; Wang, X. Prediction of refractive index of vinyl polymers by using density functional theory. *Journal of computational chemistry* **2007**, *28*, 2336–2341.

[20] Xu, J.; Chen, B.; Zhang, Q.; Guo, B. Prediction of refractive indices of linear polymers by a four-descriptor QSPR model. *Polymer* **2004**, *45*, 8651–8659.

[21] Liu, H.; Blakey, I.; Conley, W. E.; George, G.; Hill, D. J.; Whittaker, A. K. Application of quantitative structure property relationship to the design of high refractive index 193i resist. *Journal of Micro/Nanolithography, MEMS, and MOEMS* **2008**, *7*, 023001.

[22] Afzal, M. A. F.; Sonpal, A.; Haghighatlari, M.; Schultz, A. J.; Hachmann, J. A Deep Neural Network Model for Packing Density Predictions and its Application in the Study of 1.5 Million Organic Molecules. *ChemRxiv* **2019**, 8217758.

[23] Hachmann, J.; Afzal, M. A. F.; Haghighatlari, M.; Pal, Y. Building and deploying a cyberinfrastructure for the data-driven design of chemical systems and the exploration of chemical space. *Molecular Simulation* **2018**, *44*, 921–929.

[24] Afzal, M. A. F.; Vishwakarma, G.; Dudwadkar, J. A.; Haghighatlari, M.; Hachmann, J. *ChemLG* – A Program Suite for the Generation of Compound Libraries and the Survey of Chemical Space. 2019; `https://github.com/hachmannlab/chemlg`.

[25] Afzal, M. A. F.; Cheng, C.; Hachmann, J. Combining first-principles and data modeling for the accurate prediction of the refractive index of organic polymers. *Journal of Chemical Physics* **2018**, *148*.

[26] Parr, R.; Weitao, Y. *Density-Functional Theory of Atoms and Molecules*; International Series of Monographs on Chemistry; Oxford University Press, 1994.

[27] Koch, W.; Holthausen, M. C. *A chemist's guide to density functional theory*; John Wiley & Sons, 2015.

[28] Adamo, C.; Barone, V. Toward reliable density functional methods without adjustable parameters: The PBE0 model. *The Journal of Chemical Physics* **1999**, *110*, 6158–6170.

[29] Weigend, F.; Ahlrichs, R.; Peterson, K. A.; Dunning, T. H.; Pitzer, R. M.; Bergner, A. Balanced basis

sets of split valence, triple zeta valence and quadruple zeta valence quality for H to Rn: Design and assessment of accuracy. *Physical Chemistry Chemical Physics* **2005**, *7*, 3297.

[30] Von Lilienfeld, O. A. Quantum Machine Learning in Chemical Compound Space. *Angewandte Chemie International Edition* **2018**, *57*, 4164–4169.

[31] Rupp, M.; Von Lilienfeld, O. A.; Burke, K. Guest Editorial: Special Topic on Data-Enabled Theoretical Chemistry. *Journal of Chemical Physics* **2018**, *148*.

[32] DRAGON (Software for Molecular Descriptor Calculation). 2011; `http://www.talete.mi.it/`.

[33] Weininger, D. SMILES, a chemical language and information system. 1. Introduction to methodology and encoding rules. *Journal of Chemical Information and Modeling* **1988**, *28*, 31–36.

[34] Landrum, G. RDKit: Open-source cheminformatics. `http://www.rdkit.org`.

[35] Morgan, H. L. The Generation of a Unique Machine Description for Chemical Structures-A Technique Developed at Chemical Abstracts Service. *Journal of Chemical Documentation* **1965**, *5*, 107–113.

[36] Rogers, D.; Hahn, M. Extended-Connectivity Fingerprints. *Journal of Chemical Information and Modeling* **2010**, *50*, 742–754.

[37] Nilakantan, R.; Bauman, N.; Dixon, J. S.; Venkataraghavan, R. Topological torsion: a new molecular descriptor for SAR applications. Comparison with other descriptors. *Journal of Chemical Information and Computer Sciences* **1987**, *27*, 82–85.

[38] Carhart, R. E.; Smith, D. H.; Venkataraghavan, R. Atom pairs as molecular features in structure-activity studies: definition and applications. *Journal of Chemical Information and Computer Sciences* **1985**, *25*, 64–73.

[39] Schmidhuber, J. Deep learning in neural networks: An overview. *Neural Networks* **2015**, *61*, 85 – 117.

[40] Mater, A. C.; Coote, M. L. Deep Learning in Chemistry. *Journal of Chemical Information and Modeling* **2019**, *59*, 2545–2559.

[41] Vishwakarma, G. Machine Learning Model Selection for Predicting Properties of High-Refractive-Index Polymers. M.Sc. thesis, University at Buffalo, 2018.

[42] Haghighatlari, M.; Vishwakarma, G.; Altarawy, D.; Subramanian, R.; Kota, B. U.; Sonpal, A.; Setlur, S.; Hachmann, J. ChemML: A Machine Learning and Informatics Program Package for the Analysis, Mining, and Modeling of Chemical and Materials Data. *ChemRxiv* **2019**, 8323271.

[43] Haghighatlari, M.; Vishwakarma, G.; Altarawy, D.; Subramanian, R.; Kota, B. U.; Sonpal, A.; Setlur, S.; Hachmann, J. *ChemML* – A Machine Learning and Informatics Program Package for the Analysis, Mining, and Modeling of Chemical and Materials Data. 2019; `https://hachmannlab.github.io/chemml`.

[44] Girosi, F.; Jones, M.; Poggio, T. Regularization Theory and Neural Networks Architectures. *Neural Computation* **1995**, *7*, 219–269.

[45] Srivastava, N.; Hinton, G.; Alex, K.; Sutskever, I.; Salakhutdinov, R. Dropout: A Simple Way to Prevent Neural Networks from Overfittin. *Journal of Machine Learning Research 15* **2014**, *15*, 1929–1958.

[46] Caruana, R. Multitask Learning: A Knowledge-Based Source of Inductive Bias. ICML. 1993.

[47] Iovanac, N. C.; Savoie, B. M. Improved Chemical Prediction from Scarce Data Sets via Latent Space Enrichment. *The Journal of Physical Chemistry A* **2019**, *123*, 4295–4302.

[48] Nebgen, B.; Lubbers, N.; Smith, J. S.; Sifain, A. E.; Lokhov, A.; Isayev, O.; Roitberg, A. E.; Barros, K.; Tretiak, S. Transferable Dynamic Molecular Charge Assignment Using Deep Neural Networks. *Journal of Chemical Theory and Computation* **2018**, *14*, 4687–4698.

[49] Smith, J. S.; Nebgen, B. T.; Zubatyuk, R.; Lubbers, N.; Devereux, C.; Barros, K.; Tretiak, S.; Isayev, O.; Roitberg, A. E. Approaching coupled cluster accuracy with a general-purpose neural network potential through transfer learning. *Nature Communications* **2019**, *10*, 2903.

[50] Haghighatlari, M.; Shih, C.-Y.; Hachmann, J. Thinking Globally, Acting Locally: On the Issue of Training Set Imbalance and the Case for Local Machine Learning Models in Chemistry. *ChemRxiv* **2019**, 1–10.

[51] Kalervo, J.; Jaana, K. Cumulated Gain-Based Evaluation of IR Techniques. *ACM Transactions on Information Systems* **2002**, *20*, 422–446.

[52] Wang, Y.; Wang, L.; Li, Y.; He, D.; Liu, T.-Y.; Chen, W. A Theoretical Analysis of NDCG Type Ranking Measures. *Proceedings of the 26th Annual Conference on Learning Theory (COLT 2013)* **2013**, 1–30.

[53] Haghighatlari, M. Making Machine Learning Work in Chemistry: Methodological Innovation, Software Development, and Application Studies. Ph.D. thesis, University at Buffalo, 2019.

[54] Krylov, A.; Windus, T. L.; Barnes, T.; Marin-Rimoldi, E.; Nash, J. A.; Pritchard, B.; Smith, D. G.; Altarawy, D.; Saxe, P.; Clementi, C.; Crawford, T. D.; Harrison, R. J.; Jha, S.; Pande, V. S.; Head-Gordon, T. Perspective: Computational chemistry software and its advancement as illustrated through three grand challenge cases for molecular science. *Journal of Chemical Physics* **2018**, *149*, 180901.

[55] Wilkins-Diehr, N.; Crawford, T. D. NSF's inaugural software institutes: The science gateways community institute and the molecular sciences software institute. *Computing in Science & Engineering* **2018**, *20*, 26–38.