Journal of Cheminformatics

RESEARCH Open Access

Transfer learning across different chemical domains: virtual screening of organic materials with deep learning models pretrained on small molecule and chemical reaction data

Chengwei Zhang[1], Yushuang Zhai[1], Ziyang Gong[2], Hongliang Duan[3], Yuan-Bin She[1], Yun-Fang Yang[1] and An Su[1,2*]

Abstract

Machine learning is becoming a preferred method for the virtual screening of organic materials due to its cost-effectiveness over traditional computationally demanding techniques. However, the scarcity of labeled data for organic materials poses a significant challenge for training advanced machine learning models. This study showcases the potential of utilizing databases of drug-like small molecules and chemical reactions to pretrain the BERT model, enhancing its performance in the virtual screening of organic materials. By fine-tuning the BERT models with data from five virtual screening tasks, the version pretrained with the USPTO–SMILES dataset achieved R^2 scores exceeding 0.94 for three tasks and over 0.81 for two others. This performance surpasses that of models pretrained on the small molecule or organic materials databases and outperforms three traditional machine learning models trained directly on virtual screening data. The success of the USPTO–SMILES pretrained BERT model can be attributed to the diverse array of organic building blocks in the USPTO database, offering a broader exploration of the chemical space. The study further suggests that accessing a reaction database with a wider range of reactions than the USPTO could further enhance model performance. Overall, this research validates the feasibility of applying transfer learning across different chemical domains for the efficient virtual screening of organic materials.

Scientific contribution

This study verifies the feasibility of applying transfer learning to large language models in different chemical fields to help organic materials perform virtual screening. Through the comparison of transfer learning from different chemical fields to a variety of organic material molecules, the high precision virtual screening of organic materials is realized.

Keywords Organic materials, Virtual screening, Machine learning, Transfer learning, Chemical reaction database

*Correspondence:
An Su
ansu@zjut.edu.cn
Full list of author information is available at the end of the article

Introduction

Organic materials, such as organic photovoltaics (OPVs), organic light-emitting diodes (OLEDs), and organic redox flow batteries (ORFBs), play a crucial role in materials science and their unique functional properties are widely utilized in various research fields [1–4]. Since organic materials have complex structures, it is often costly to explore them through wet experiments. Therefore, virtual screening helps to screen target organic materials before conducting wet experiments. Starting from the simplest Quantitative Structure–Activity Relationships (QSAR) models, virtual screening has a long history in the field of drug discovery [5–8]. Nowadays, in the era of artificial intelligence, different types of state-of-the-art (SOTA) machine learning model architectures have been developed for the virtual screening of drug-like small molecules, including Transformers [9], graph neural networks [10, 11], sequence-graph hybrid models [12], and large language models (LLMs) [13].

On the other hand, in 2015, the Aspuru-Guzik group proposed the concept of a “computational funnel” to explain the virtual screening procedure of organic materials [14]. The process comprises several levels, each increasing in computational intensity. Initially, machine learning techniques rapidly predict material properties (e.g., HOMO and LUMO positions, optical properties) by exploiting complex relationships within relevant chemical subsets, efficiently identifying promising candidates with minimal resource expenditure. Later in the funnel, Density Functional Theory (DFT) computations provide detailed and accurate analyses of the remaining candidates. The process is finalized by an experimental test, which is slow and costly, but the number of candidates reaching this stage has been minimized by the previous levels [14]. While several studies utilized machine learning methods for the property prediction of organic materials [15–18], they could not avoid one major challenge—the limited number of training data.

There are several ways to address the problem of scarce training data. Some researchers have chosen to augment the dataset by supplying a wider range of physical and chemical information at the molecular level [19–21]. Another common strategy is transfer learning [18, 22, 23], a popular machine learning technique that involves pre-training with a large dataset and fine-tuning using smaller datasets, which allows for efficient learning using limited resources. However, traditional transfer learning based on supervised learning requires similar types of targets in the pre-training dataset and fine-tuning datasets. Otherwise, the transfer learning may produce results opposite to those expected [24]. On the other hand, recent studies suggest that the Bidirectional Encoder Representations from Transformers (BERT) model [25] with an unsupervised pre-training phase and a supervised fine-tuning phase may be able to address the limitation of supervised transfer learning in chemistry. Schwaller et al. developed and successfully used a BERT-based framework, *rxnfp*, for predictive chemical reaction classification, which pre-trained and fine-tuned the BERT models using databases consisting of different classification systems [26]. In our group’s previous studies on PorphyBERT [17] and SolvBERT [27], we found that unsupervised learning using BERT models and their subsequent fine-tuning allows the model to learn chemistry from a large number of molecules without being affected by differences in the targets of the pre-training dataset and fine-tuning datasets.

Since the pre-training phase of the BERT model involves unsupervised learning and does not require any property or activity data of the molecules, it is theoretically possible to incorporate a large number of molecular structures into the pre-training process, thereby generating a pre-trained model with a wealth of knowledge in a chemical space larger than one with only organic materials. Therefore, in addition to the organic materials database, drug-like small molecule data can be included in the pre-training data, which could potentially broaden the model’s horizon for the variety of organic parts of the organic materials. Meanwhile, chemical reaction databases can offer a wider range of chemical structures, including organic and inorganic materials, metals, complexes, and molecular associations. Chemical reaction data have been extensively studied for pretraining models to predict reaction-specific properties such as reaction outcomes [22, 28, 29], classifications [26], and yields [30]. However, to our knowledge, they have not yet been explored for predicting molecular properties.

In this study, we first pretrained the BERT models using large databases such as chemical reaction data, drug-like small molecule data, and organic materials data. After pretraining, we fine-tuned these models using small organic materials databases for different prediction tasks. In addition, to further explain the performance of the models trained on different combinations of the pretraining and fine-tuning databases, we summarized their statistics of organic build blocks and visualized their chemical space. Furthermore, we investigated the effect of the size of the pretraining and fine-tuning datasets on the predictive performance of the models. Finally, we compared the best models proposed in this study with the models without an unsupervised pretraining phase and with the Transformer/BERT models pretrained using other databases.

Methods

Datasets for pretraining

ChEMBL

ChEMBL is a manually curated database of bioactive molecules with drug-like properties. It brings together chemical, bioactivity, and genomic data to help translate genomic information into effective new drugs [31]. The data used in this study were obtained from the ChEMBL download channel (https://www.ebi.ac.uk/chembl, accessed April 10, 2023), which contains Simplified Molecular Input Line Entry System (SMILES) [32] data for 2,327,928 drug-like small molecules.

Clean energy project database (CEPDB)

In 2008, Aspuru-Guzik et al. launched the Harvard Clean Energy Project (CEP) to help find high-efficiency organic photovoltaic materials [33]. They built the main CEP using a combinatorial molecular generator, and by 2019, the project team had synthesized at least 2,322,849 molecules. In this study, CEPDB data was downloaded from https://figshare.com/articles/dataset/moldata_csv/9640427 (accessed April 12, 2023), from which 10^4, 10^5, and 10^6 molecules were randomly selected as training datasets, referred to as CEPDB-10K, CEPDB-100K, and CEPDB-1M, respectively.

United States patent and trademark office (USPTO) databases

The USPTO database (https://figshare.com/articles/dataset/Chemical_reactions_from_US_patents_1976-Sep2016_/5104873, accessed on April 13, 2023) contains reactions extracted through text-mining from U.S. patents published between 1976 and September 2016, available as reaction SMILES. The USPTO database used in this study was derived from the study by Giorgio Pesciullesi et al. [22] and contains 1,048,575 reactions. In addition, the molecules were extracted from these chemical reactions, resulting in 5,390,894 molecules that comprised the USPTO–SMILES dataset. Furthermore, duplicate molecules were removed from the molecules in the USPTO–SMILES dataset to obtain the USPTO–SMILES-clean dataset, which contains 1,345,854 molecules with different SMILES.

Datasets for fine-tuning and evaluation

Metalloporphyrin and porphyrin database (MpDB)

Porphyrins are a class of heterocyclic macrocycle organic compounds containing four modified pyrrole substituents interconnected by methyl bridges (=CH–) at their α-carbon atoms. Porphyrins are the active central structure of chlorophyll and have tunable photochemical properties when coordinated with metal ions. MpDB is derived from the Computational Materials Repository database (CMR) of porphyrin-based dyes database [34, 35] (https://cmr.fysik.dtu.dk/dssc/dssc.html, accessed on March 23, 2023). MpDB contains 12,096 porphyrins or metalloporphyrins, including detailed structural and energy level information. The porphyrin dye structures in MpDB were converted to canonical SMILES using our previously developed framework [17]. This database has been utilized to train machine learning models for predicting HOMO–LUMO gaps [17, 18, 36], which also served as a virtual screening task in our study.

Benzodithiophene organic photovoltaics (OPV–BDT)

Organic Photovoltaics (OPVs) are solar cells that utilize organic materials, primarily carbon-based compounds, to convert sunlight into electricity. In 2019, St. John et al. created a database of 91,000 candidate molecules for OPV applications [37], which was subsequently employed for machine learning to predict the properties of OPVs [37, 38]. This database includes various photoelectric properties of the molecules, including HOMO energy, LUMO energy, HUMO–LUMO gap, and spectral overlap. In our study, we focused on a subset of 10,248 OPVs that contain benzodithiophene (BDT), a typical donor material in OPVs, and renamed this subset as OPV–BDT. A virtual screening task was then conducted to predict the HOMO–LUMO gap of the OPV–BDT molecules.

Experimental database of optical properties of organic compounds (EOO)

EOO is a comprehensive collection of data on the optical characteristics of organic chromophores—an essential class of compounds in photochemistry. Initially created by Joung et al. in 2020, this database contains 20,236 data points, encompassing 7016 unique organic chromophores in solvents or solid states [39]. The EOO database has previously been employed to train a graph convolutional network (GCN) for predicting the maximum absorption wavelength (MAW) and maximum emission wavelength (MEW) of these compounds [13, 40, 41]. For our study, we focused on the same two predictive tasks, utilizing 17,295 data points for MAW and 18,142 data points for MEW.

Database of organic donor–acceptor molecules (solar)

Organic donor–acceptor molecules are an important class of compounds in organic photochemistry due to their unique electronic and optical properties. A database of organic donor–acceptor molecules is publicly available in the CMR (https://cmr.fysik.dtu.dk/solar/solar.html, accessed on March 23, 2023), which contains the Kohn–Sham (B3LYP) HOMO, LUMO, HOMO–LUMO gap, and singlet–triplet gap of organic donor–acceptor molecules. We selected the Kohn–Sham HOMO–LUMO gap (KS_gap) as the label for the virtual screening task, with a total of 5366 data.

The distributions of the property values in all the fine-tuning datasets introduced above are shown in Figure S1.

Models

BERT

Bidirectional Encoder Representations from Transformers (BERT) is a pretrained natural language processing model developed by Google [25]. BERT has two distinct phases: pretraining and fine-tuning. During pretraining, BERT learns about the contextual relationships between words and sentences through unsupervised learning on large amounts of text data. Fine-tuning involves additional training by virtual screening classification or regression tasks. In this way, the pretrained model is better able to predict the specific nuances of the virtual screening task, ultimately achieving higher performance. Fine-tuning BERT typically requires the use of only a small amount of task-specific data and is therefore much faster than pretraining.

The BERT model architecture used in this study was built based on the *rxnfp* framework [26]. The pretraining data was randomly divided into two parts: training set and validation set, with a ratio of 9:1 and training epochs of 20. During pretraining, only the SMILES of molecules were provided to the model, and the model learned the structural information of molecules in an unsupervised learning way. For fine-tuning, the regression task dataset used for the fine-tuning phase was randomly split into training, validation, and test sets in a fine-tuning ratio of 8:1:1. Each fine-tuning was done with 50 training epochs and the learning rate was set to 10^{-5}. The most proficient pretrained model underwent further optimization in the fine-tuning stage, with specific adjustments made to the learning rate and the dropout probability of neurons within the hidden layers.

Baseline models

Two classical machine learning models, Random Forest (RF) and Support Vector Machine Regression (SVR) were used as baseline models. The molecular representation selected for these two models is MAP4, a molecular fingerprint that combines circular substructures and atom pairs [42]. MAP4 offers a broader range of applicability compared to other fingerprints developed specifically for drug-like small molecules [42]. It has been utilized in data-driven synthesis planning and property prediction [43, 44]. In addition, a graph convolutional network (GCN) developed by Shen et al. [44] was chosen as another baseline model. GCN reads the SMILES of a molecule and converts the molecule into a graph format. Each of the three baseline models was trained directly by the regression tasks on organic materials in a supervised learning way. Furthermore, DeepChem-77M is a pretrained model based on the RoBERTa [45] transformer using 77M unique SMILES from ChemBERTa [46] for pretraining. In this study, the model was directly fine-tuned by the regression tasks on organic materials.

Chemical space visualization

The SMILES of the molecule are partially encoded by the unsupervised learning model during the pre-training process, generating a molecular fingerprint. The dimensionality of the fingerprint was reduced by Tree Map (TMAP), where the high-dimensional data was transformed into a tree structure through the MinHash algorithm [47]. Finally, the Molecular fingerprints with reduced dimensions were visualized by an interactive visualization tool called Faerun [48], displaying both the local and global data structure of the chemical space.

Results and discussions

In this study, we adopted BERT, one of the SOTA models in natural language processing, as the foundational model architecture. The training of BERT was divided into a pretraining and a fine-tuning phase, as shown in Fig. 1. The BERT models were pretrained using databases in different chemical domains, such as USPTO (chemical reactions), ChEMBL (drug-like small molecules), and CEPDB (organic materials), respectively. The pretrained models were then fine-tuned for different regression tasks in organic materials, including the band gap in MpDB and OPV–BDT, KS_gap in Solar, and the MAW and MEW in EOO. It is worth noting that these five virtual screening tasks differ significantly in terms of the range of values and type of properties. The major hyperparameters for the pretraining and fine-tuning phases are listed in Tables S1 and S2.

To eliminate the possible effects of differences in the structure of Reaction SMILES (the original format of USPTO) and normal SMILES data, we further prepared two additional USPTO datasets by extracting the SMILES of all the reaction components from the USPTO dataset, generating USPTO–SMILES (Fig. 2, step 1), and then further removed the duplicate SMILES in USPTO–SMILES to generate USPTO–SMILES-clean (Fig. 2, step 2).

Performance of pretrained BERT models for virtual screening tasks

Figure 3 illustrates the prediction performance of the five models pretrained using different databases and subsequently fine-tuned with virtual screening tasks. Given the challenge of directly comparing absolute errors across tasks due to variations in their units, the evaluation

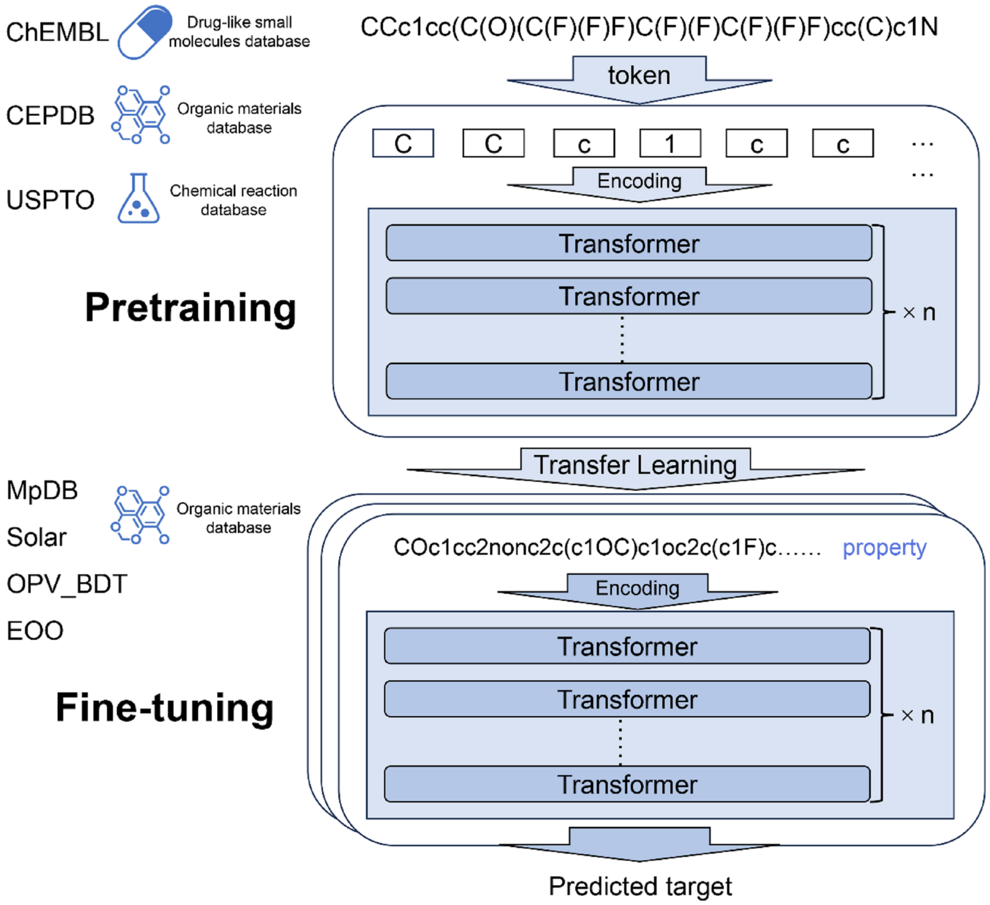

Fig. 1 The workflow of this study. The models were pretrained using chemical databases from different chemical domains and then fine-tuned individually for different virtual screening tasks in organic materials

primarily utilizes the R^2 metric, with detailed results on Mean Absolute Error (MAE) and Root Mean Square Error (RMSE) for each model provided in Tables S3 to S5. The standard deviation and significance test are provided in Table S6. Among the evaluated models, those pretrained on the USPTO–SMILES database consistently achieved the highest performance across all five virtual screening tasks. Intriguingly, the model pretrained using the CEPDB database, which is the sole database dedicated to organic materials, exhibited the lowest performance in four out of the five tasks. This outcome underscores the significant impact of the choice of the pretraining database on the model's effectiveness in virtual screening applications.

Figure 3 illustrates that among the three models trained using differently processed USPTO data, those pretrained on USPTO–SMILES consistently outperform the model using raw USPTO data. This improvement highlights the benefits of extracting molecular SMILES from the raw data, which likely enhances pretraining by eliminating non-structural information (Fig. 2). Moreover, the USPTO–SMILES pretrained model also surpasses the performance of the model pretrained on USPTO–SMILES-clean, despite both datasets sharing identical data structures and covering similar chemical spaces. This superiority can be attributed to the USPTO–SMILES dataset being four times larger than its clean counterpart. The repetition of SMILES in the USPTO–SMILES dataset often represents molecules that are more commonly encountered in chemical reactions. This repetition aids in a deeper understanding of chemical language, suggesting that a larger, more repetitive dataset can be beneficial for model training in this context.

However, the size of the pretraining database should be carefully controlled. We pretrained another model using all the molecules from the USPTO–SMILES, ChEMBL, and CEPDB databases, referred to as "SMILES-all" in Fig. 3. Unexpectedly, the performance of this model was worse than that of the models pretrained on any of the individual databases. This phenomenon suggests that an excessive amount of data may not necessarily improve prediction performance and can, in fact, be detrimental.

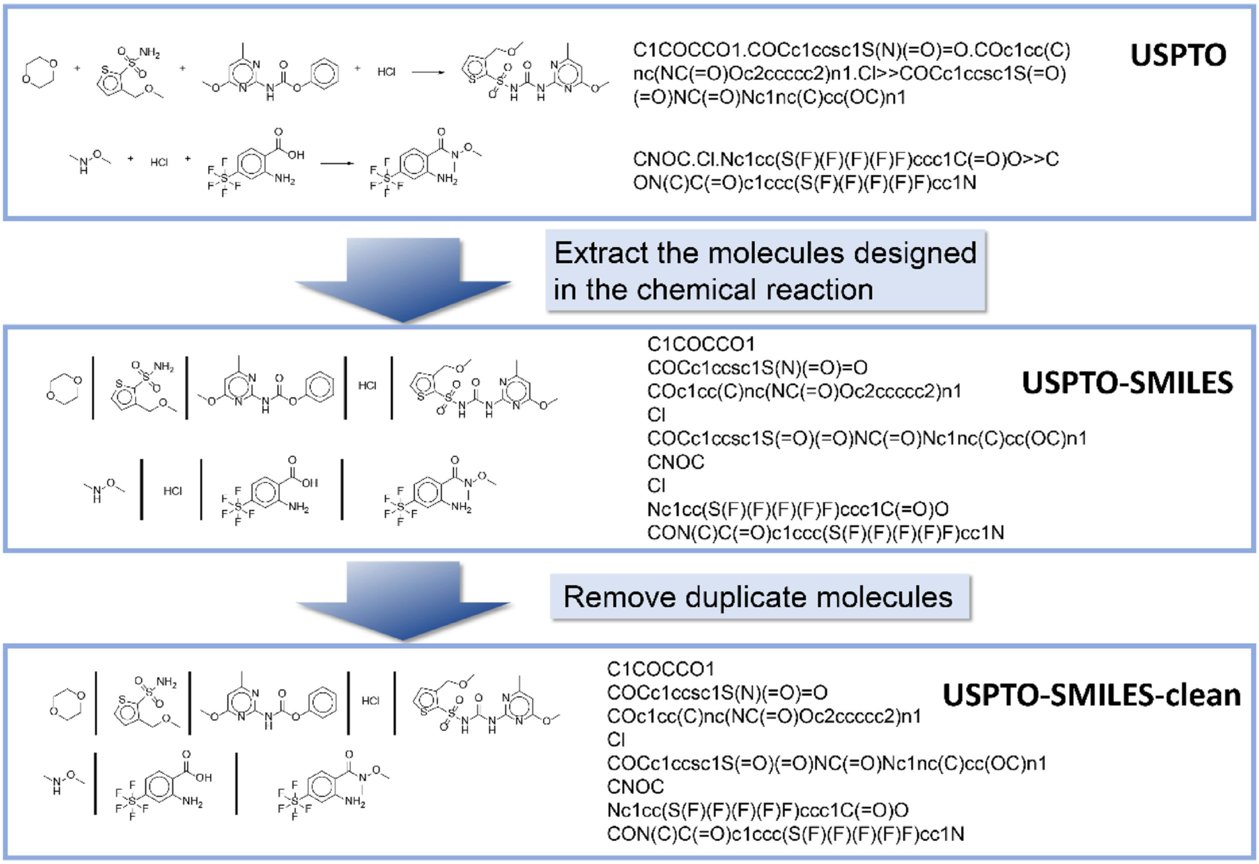
USPTO
C1COCCO1.COCc1ccsc1S(N)(=O)=O.COc1cc(C)nc(NC(=O)Oc2ccccc2)n1.Cl>>COCc1ccsc1S(=O)(=O)NC(=O)Nc1nc(C)cc(OC)n1
CNOC.Cl.Nc1cc(S(F)(F)(F)(F)F)ccc1C(=O)O>>CON(C)C(=O)c1ccc(S(F)(F)(F)(F)F)cc1N
Extract the molecules designed in the chemical reaction
USPTO-SMILES
C1COCCO1
COCc1ccsc1S(N)(=O)=O
COc1cc(C)nc(NC(=O)Oc2ccccc2)n1
Cl
COCc1ccsc1S(=O)(=O)NC(=O)Nc1nc(C)cc(OC)n1
CNOC
Cl
Nc1cc(S(F)(F)(F)(F)F)ccc1C(=O)O
CON(C)C(=O)c1ccc(S(F)(F)(F)(F)F)cc1N
Remove duplicate molecules
USPTO-SMILES-clean
C1COCCO1
COCc1ccsc1S(N)(=O)=O
COc1cc(C)nc(NC(=O)Oc2ccccc2)n1
Cl
COCc1ccsc1S(=O)(=O)NC(=O)Nc1nc(C)cc(OC)n1
CNOC
Nc1cc(S(F)(F)(F)(F)F)ccc1C(=O)O
CON(C)C(=O)c1ccc(S(F)(F)(F)(F)F)cc1N
HCl

Fig. 2 The workflow for creating the USPTO–SMILES and USPTO–SMILES-clean databases. The SMILES of molecules in USPTO were extracted to create USPTO–SMILES, while the duplicate SMILES in USPTO–SMILES were removed to create USPTO–SMILES-clean

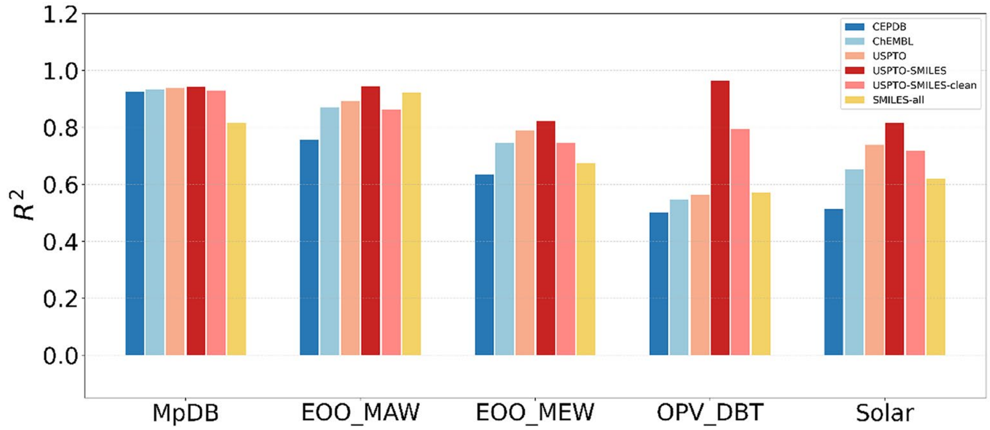

Fig. 3 R^2 values of transfer learning using different databases for pretraining

A control experiment was conducted by comparing the performance of the USPTO–SMILES pretrained model, fine-tuned with the MpDB and EOO_MAW virtual screening tasks, against the same model architecture trained directly on these tasks without USPTO–SMILES pretraining. The R^2 of the directly trained models was 0.048 lower than that of the pre-trained model for MpDB and 0.166 lower for EOO_MAW (Table S3).

Statistics of organic building blocks in the databases

The superior performance of models pretrained with the USPTO–SMILES and ChEMBL databases over those using the CEPDB, an organic material database, can be attributed to the richer diversity of organic building blocks available in the former two (Fig. 4). Utilizing RDKit's Chem.Fragments package, we quantified the number of common organic building blocks across the pretraining and fine-tuning databases, with these findings detailed in Fig. 5 and further elaborated in Table S7. Figure 5a–c reveal that USPTO–SMILES and ChEMBL contain a significantly broader spectrum of organic building blocks compared to CEPDB. Specifically, of the 85 identified organic building blocks, CEPDB lacked 72, whereas USPTO–SMILES and ChEMBL were missing only one. This extensive repository of organic building blocks in USPTO–SMILES and ChEMBL is crucial for the enhanced performance of the BERT models pretrained on these databases for the virtual screening tasks.

Chemical space coverage

We aimed to explore the extent of chemical space encompassed by the databases used for pretraining and fine-tuning our models. To achieve this, we applied TMAP [47] for dimensionality reduction on the fingerprints generated by the USPTO–SMILES model and visualized the resultant chemical space using Faerun [48]. As depicted in Fig. 6, the chemical space covered by USPTO–SMILES (highlighted in orange) is broader than that covered by both ChEMBL (in green) and CEPDB (in blue). This broader coverage by USPTO–SMILES can be more comprehensively explored through the interactive TMAP visualizations available in the S.I. The expansive chemical space coverage of USPTO–SMILES accounts for its pretrained models outperforming those pretrained on ChEMBL, despite both databases featuring a similar diversity of organic building blocks. This superior performance can be attributed to USPTO–SMILES not only encompassing drug-like small molecules found in ChEMBL but also including metals, non-metallic inorganic compounds, and organic materials frequently used as catalysts or reagents in reactions. In addition, since TMAP is an unsupervised learning technique, molecules with similar structures tend to be adjacent to each other on the TMAP, even if they are not from the same database.

Effect of the relative size of pretraining and fine-tuning datasets

To better comprehend the impact of pretraining and fine-tuning dataset sizes on prediction performance, we randomly selected 2 million to 10,000 data instances from USPTO–SMILES and 1 million to 10,000 data instances from ChEMBL and CEPDB, respectively, and pretrained the models using these sampled datasets (Fig. 7). Within the framework of the USPTO–SMILES and ChEMBL

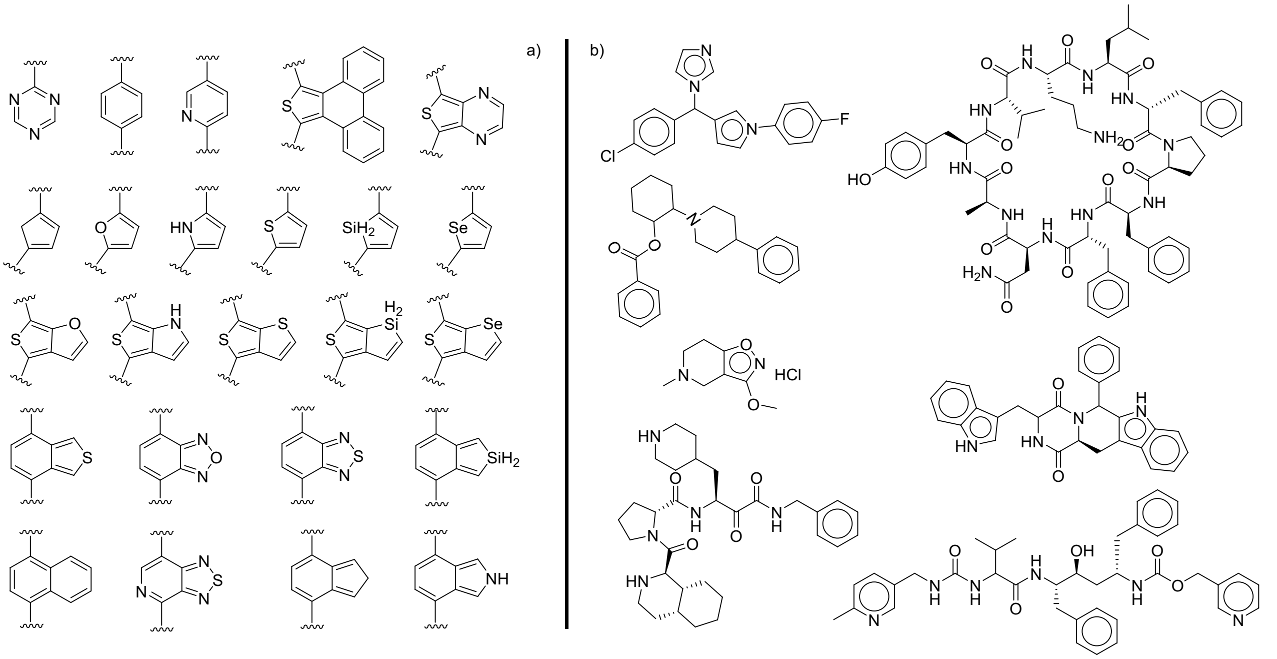

Fig. 4 **a** The 26 building blocks used for generating the CEPDB; **b** molecules in the ChEMBL database are displayed, selected at random from the database

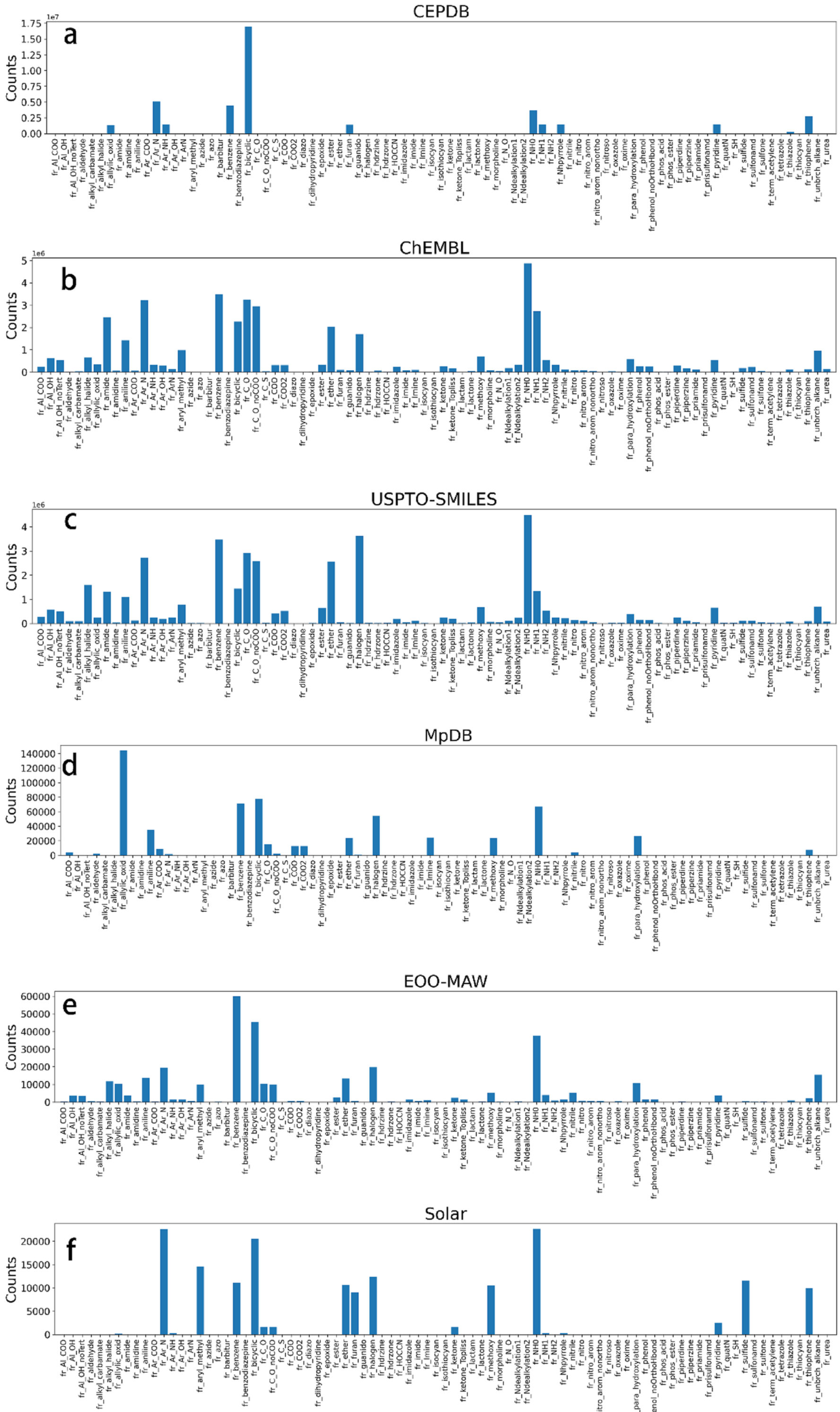

Fig. 5 Statistics of organic building blocks in **a** CEPDB, **b** ChEMBL, **c** USPTO–SMILES, **d** MpDB, **e** EOO_MAW, and **f** solar. For the details of these organic building blocks, an alternative tabular form of this summary is shown in Table S7. Detailed information about the statistical structure is in Figure S2

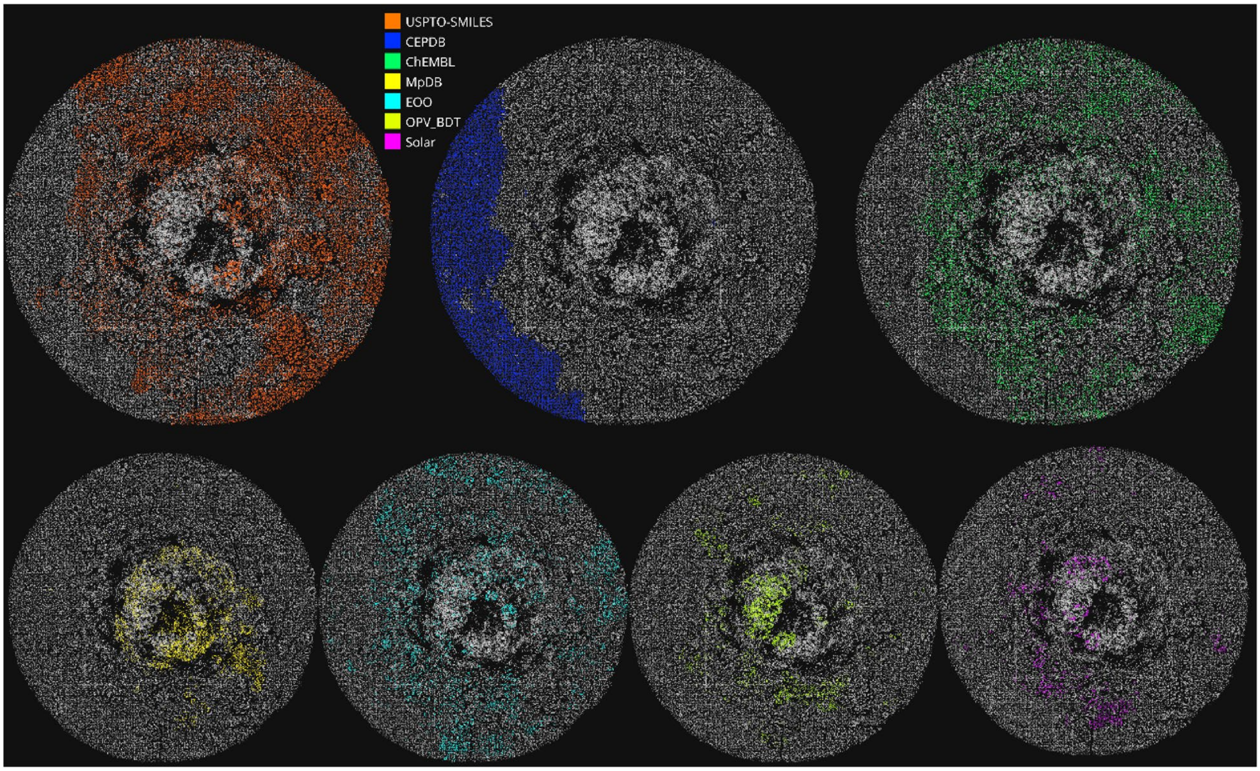

Fig. 6 The two-dimensional chemical space of USPTO–SMILES, CEPDB, ChEMBL, Solar, MpDB, OPV_BDT, and EOO. The molecules representation, dimensionality reduction, and visualization were implemented by BERT, TMAP [47], and Faerun [48], respectively. To maximize the visualization, 50,000, 20,000, and 20,000 data were randomly selected from USPTO–SMILES, ChEMBL, and CEPDB, respectively, for visualization

pre-training datasets, the R^2 for four out of the five tasks assessed typically demonstrated a decreasing trend with the reduction in the size of the pre-training dataset. Conversely, this decline was not evident in the tasks pre-trained on the CEPDB dataset, likely a consequence of the previously discussed insufficiency in chemical information inherent to CEPDB.

Comparison with the baseline models

To further understand the predictive capability of USPTO-pretrained models for the virtual screening of organic materials, the best-performed USPTO-pretrained models in this study were compared with four baseline models, encompassing two traditional machine learning models, a GCN [44], and the DeepChem-77M [46, 49] model trained on a Roberta architecture (Table 1). The hyperparameters of these models are provided in Tables S8 and S9. The findings indicate that our USPTO–SMILES model surpasses all competitors across four of the five evaluated tasks, underscoring the superiority of our approach in virtual screening applications. Furthermore, the comparative underperformance of DeepChem-77M could be ascribed to its pre-training on the PubChem database, which offers a narrower chemical space than USPTO as discussed above, and its adoption of a Roberta architecture with fewer hidden layers than BERT, potentially diminishing its performance.

In addition, different fractions of the OPV–BDT dataset were used to demonstrate the effect of data size on the performance of our USPTO–SMILES model compared to the RF baseline model. As shown in Fig. 8, while the RF performs better with less than 40% of the OPV–BDT dataset and both models have R^2 values lower than 0.85, the USPTO–SMILES model surpasses the RF when the training data fraction exceeds 40%. Moreover, the performance gap between the two models widens as the training data size increases. This is because BERT models are more complex than simple machine learning models and require more data to learn and generalize patterns effectively. Once a certain data threshold is reached, their generalization ability exceeds traditional models, leading to superior performance. In the context of virtual screening of organic materials, obtaining more training data and improving the predictive performance of models can synergistically enhance the accuracy of virtual screening. It is advisable to acquire additional training data (e.g., through DFT computation) and develop a higher-performing model (i.e., $R^2 > 0.95$) rather than

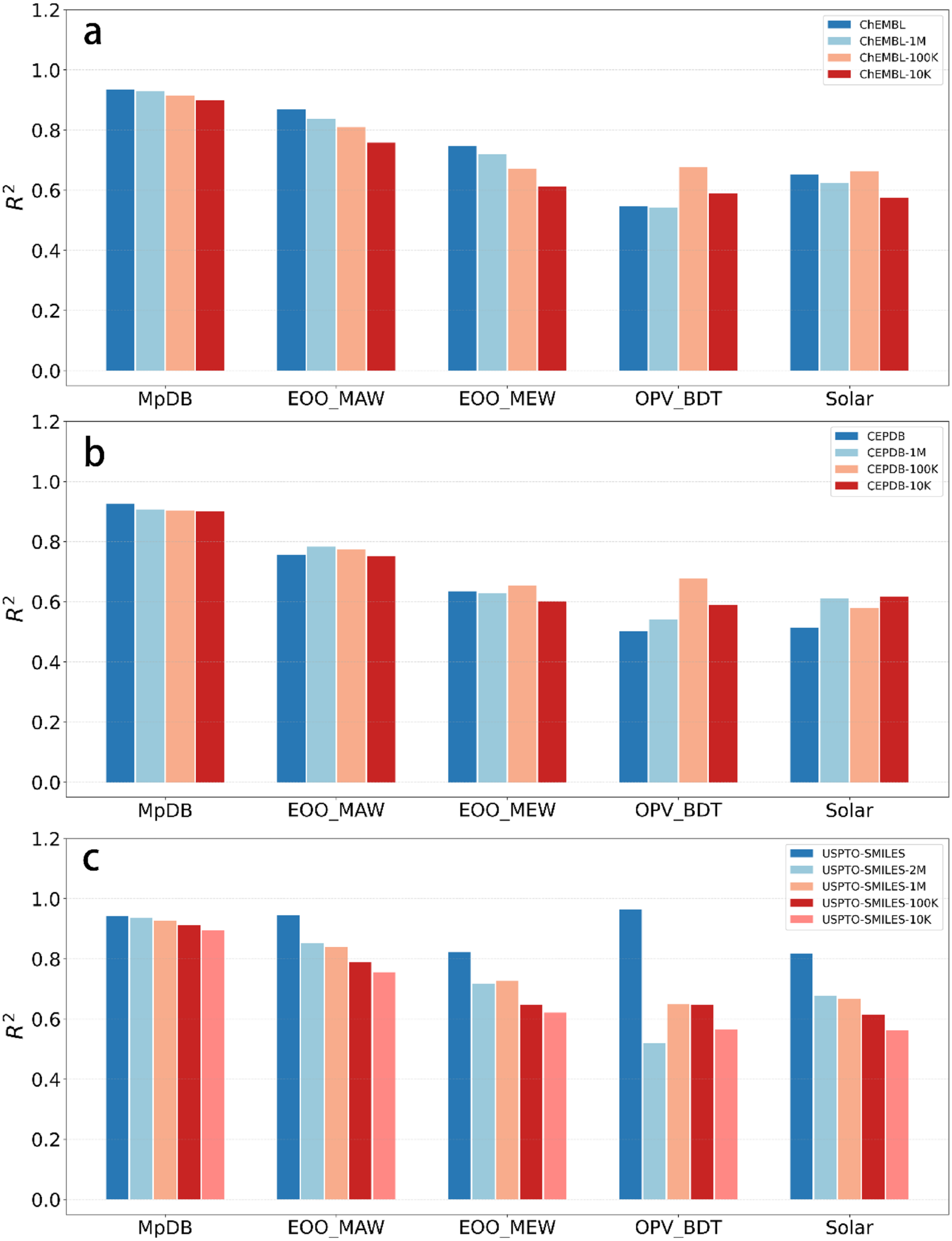

Fig. 7 R^2 value of five virtual screening prediction tasks with models pretrained using random samples of **a** CEPDB, **b** ChEMBL, and **c** USPTO–SMILES of different sizes

Table 1 The performance comparison between the USPTO–SMILES model and the baseline models

R^2	MpDB	EOO_MAW	EOO_MEW	OPV_BDT	Solar
USPTO–SMILES (our study)	**0.9428**	**0.9443**	**0.8229**	**0.9637**	0.8174
Support vector regressor (SVR)	0.9320	0.2086	0.2187	0.0862	0.7714
Random forest (RF)	0.8953	0.8959	0.8081	0.8748	**0.8416**
GCN [44]	0.8950	0.8384	0.6809	0.8827	0.8045
DeepChem-77M [49]	0.9097	0.8098	0.6457	0.5750	0.6158

The bold values indicate the maximum values in each column

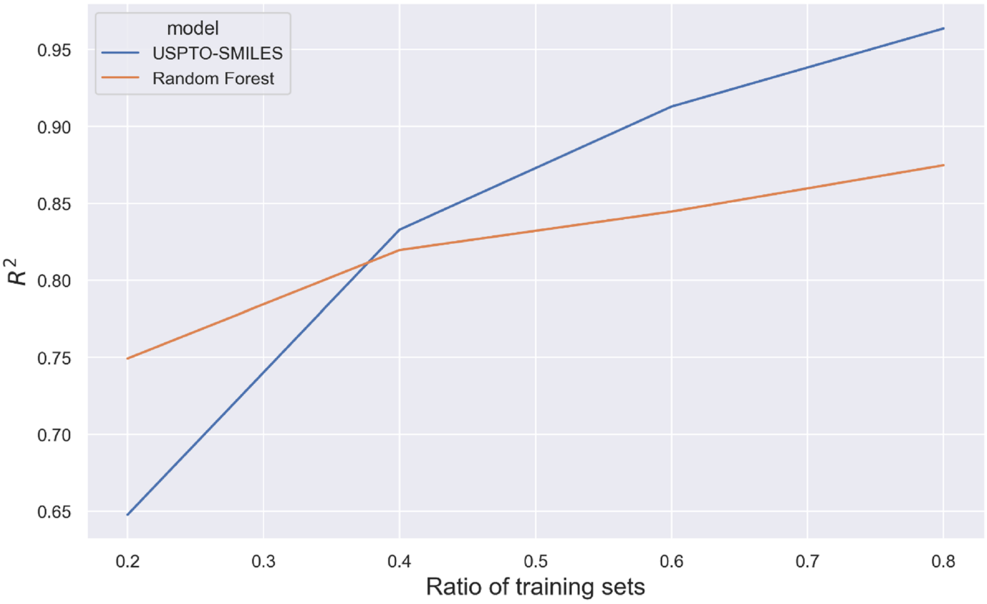

Fig. 8 The impact of different fractions of the OPV–BDT dataset on the prediction performance of the USPTO–SMILES and the RF models

Table 2 Computing time consumed by various models

Cost time (min)	MpDB	EOO_MAW	EOO_MEW	OPV_BDT	Solar
USPTO–SMILES	40.3	58.1	60.5	35.7	37.8
SVR	10.6	21.4	25.4	6.6	3.2
RF	9.3	12.8	11.8	6.7	10.9
GCN	23.1	15.8	31.2	35.4	28.7
Deepchem-77M	9.8	15.5	23.1	28.2	18.5

SVR and RF run on Intel Core i7-11700, and other models run on NVIDIA GeForce RTX 3060

settling for a mediocre model (i.e., $R^2 < 0.85$) without further computation.

Furthermore, the time required for the complete training of these models is shown in Table 2. Although the training of our USPTO–SMILES model took longer than other machine learning-based and graph neural network-based models, all training times were roughly within 1 h. Given that these training sessions were conducted on a PC equipped with an affordable NVIDIA GeForce RTX 3060 GPU, the efficiency of our model is considered acceptable for daily research purposes.

Our approach to enhancing virtual screening methodologies could see significant improvements by incorporating a commercial database like Pistachio, which offers a more comprehensive collection of reaction data than the USPTO by including reactions from the European Patent Office (EPO). Access to Pistachio directly is, unfortunately, not possible for our team. However, we utilized the open-source rxnfp-Pistachio, a model pretrained on Pistachio by Schwaller et al. [26] (https://rxn4chemistry.github.io/rxnfp/, accessed on March 29, 2023), for preliminary comparisons. According to our initial analysis presented in Tables S3 to S5, replacing the USPTO database with Pistachio for pretraining could enhance the performance of virtual screening, especially in the EOO_MEW and Solar tasks. It is important to clarify that this observation does not detract from our original findings. Rather, it reinforces our proof-of-concept that leveraging chemical reaction data is advantageous for the virtual screening of organic materials. We advocate for further investigation into the potential of the Pistachio database by researchers who have access to it.

Additionally, a rough comparison was conducted between our USPTO–SMILES model and other models from previous studies using the same data source (Table 3). Our model outperformed the Transformer model on the MpDB dataset and the GCN model on the EOO–MAW dataset. Although the ASGN model achieved a lower MAE than our model, it was trained on the entire OPV database, whereas our model was trained on a subset containing only BDT donors.

Conclusions

In this study, we have successfully demonstrated the concept of pretraining deep learning models on databases not specifically related to organic materials, such as ChEMBL and USPTO, for the virtual screening of organic materials. Leveraging the unique unsupervised pretraining phase of the BERT model, we demonstrate the feasibility of transfer learning across diverse chemical domains, including organic materials, drug-like small molecules, and chemical reactions. Our findings reveal that among the various BERT models pretrained on different databases, those pretrained using molecular SMILES data extracted from the USPTO database exhibited superior predictive performance in the majority of virtual screening tasks. This enhanced performance can be attributed to the USPTO database's broader variety of organic building blocks and its more extensive coverage of chemical space. Our study underscores the potential of applying cross-domain transfer learning to address the challenge of data scarcity in the virtual screening of organic materials and possibly other chemical categories. By showcasing the effectiveness of pretraining deep learning models on diverse chemical databases, we aim to inspire further research in this direction, encouraging the exploration of more extensive databases like Pistachio and fostering advancements in the field of virtual screening.

Table 3 Comparison with the models developed in previous studies

Model	R^2	RMSE	MAE
MpDB			
Our work	0.9429	0.1064	0.0931
Transformer [17]	0.8010	0.1965	0.1524
OPV			
Our work (trained on BDT subset)	0.9637	0.0775	0.0530
ASGN [38] (trained on full dataset)			0.028
EOO–MAW			
Our work	0.9428	25.6324	15.0124
GCN [40]		31.6	

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s13321-024-00886-1.

Supplementary Material 1.

Acknowledgements
This research was supported by the National Natural Science Foundation of China under Grant No. 22108252, the Joint Funds of the Zhejiang Provincial Natural Science Foundation of China under Grant No. LHDMZ23B060001, and Zhejiang Provincial Key R&D Project under Grant No. 2022C01179.

Author contributions

AS conceptualized the study. CZ conducted the experiments and wrote the main manuscript text. YZ and ZG prepared the figures and datasets. HD and AS revised the main manuscript text. YBS and YFY supervised the project and provided resources. All authors read and approved the final manuscript.

Data availability

The datasets and code supporting the conclusions of this article are available in GitHub (https://github.com/su-group/Virtual-Screening-of-Organic-Materials.git) and archived in the figshare repository (https://doi.org/10.6084/m9.figshare.24679305).

Declarations

Competing interests

The authors declare no competing interests.

Author details
[1]State Key Laboratory Breeding Base of Green Chemistry-Synthesis Technology, Key Laboratory of Green Chemistry-Synthesis Technology of Zhejiang Province, College of Chemical Engineering, Zhejiang University of Technology, Hangzhou 310014, Zhejiang, China. [2]Key Laboratory of Pharmaceutical Engineering of Zhejiang Province, Key Laboratory for Green Pharmaceutical Technologies and Related Equipment of Ministry of Education, Collaborative Innovation Center of Yangtze River Delta Region Green Pharmaceuticals, Zhejiang University of Technology, Hangzhou 310014, People's Republic of China. [3]Faculty of Applied Sciences, Macao Polytechnic University, Macao 999078, China.

Received: 5 March 2024 Accepted: 21 July 2024
Published online: 30 July 2024

References

1. Ostroverkhova O (2016) Organic optoelectronic materials: mechanisms and applications. Chem Rev 116(22):13279–13412
2. Hedley GJ, Ruseckas A, Samuel IDW (2017) Light harvesting for organic photovoltaics. Chem Rev 117(2):796–837
3. Zou S-J, Shen Y, Xie F-M, Chen J-D, Li Y-Q, Tang J-X (2020) Recent advances in organic light-emitting diodes: toward smart lighting and displays. Mater Chem Front 4(3):788–820
4. Luo J, Hu B, Hu M, Zhao Y, Liu TL (2019) Status and prospects of organic redox flow batteries toward sustainable energy storage. ACS Energy Lett 4(9):2220–2240
5. Stokes JM, Yang K, Swanson K, Jin W, Cubillos-Ruiz A, Donghia NM, MacNair CR, French S, Carfrae LA, Bloom-Ackermann Z, Tran VM, Chiappino-Pepe A, Badran AH, Andrews IW, Chory EJ, Church GM, Brown ED, Jaakkola TS, Barzilay R, Collins JJ (2020) A deep learning approach to antibiotic discovery. Cell 181(2):475–483
6. Li H, Sze K-H, Lu G, Ballester PJ (2021) Machine-learning scoring functions for structure-based virtual screening. WIREs Comput Mol Sci 11(1):e1478
7. Wieder O, Kohlbacher S, Kuenemann M, Garon A, Ducrot P, Seidel T, Langer T (2020) A compact review of molecular property prediction with graph neural networks. Drug Discov Today Technol 37:1–12

8. Muratov EN, Bajorath J, Sheridan RP, Tetko IV, Filimonov D, Poroikov V, Oprea TI, Baskin II, Varnek A, Roitberg A, Isayev O, Curtalolo S, Fourches D, Cohen Y, Aspuru-Guzik A, Winkler DA, Agrafiotis D, Cherkasov A, Tropsha A (2020) QSAR without borders. Chem Soc Rev 49(11):3525–3564
9. Schwaller P, Laino T, Gaudin T, Bolgar P, Hunter CA, Bekas C, Lee AA (2019) Molecular transformer: a model for uncertainty-calibrated chemical reaction prediction. ACS Cent Sci 5(9):1572–1583
10. Yang K, Swanson K, Jin W, Coley C, Eiden P, Gao H, Guzman-Perez A, Hopper T, Kelley B, Mathea M (2019) Analyzing learned molecular representations for property prediction. J Chem Inf Model 59(8):3370–3388
11. Wang Y, Wang J, Cao Z, Barati Farimani A (2022) Molecular contrastive learning of representations via graph neural networks. Nat Mach Intell 4(3):279–287
12. Ying C, Cai T, Luo S, Zheng S, Ke G, He D, Shen Y, Liu T-Y (2021) Do transformers really perform badly for graph representation? Adv Neural Inf Process Syst 34:28877–28888
13. Boiko DA, MacKnight R, Kline B, Gomes G (2023) Autonomous chemical research with large language models. Nature 624(7992):570–578
14. Pyzer-Knapp EO, Suh C, Gómez-Bombarelli R, Aguilera-Iparraguirre J, Aspuru-Guzik A (2015) What is high-throughput virtual screening? A perspective from organic materials discovery. Annu Rev Mater Res 45(1):195–216
15. Wen Y, Fu L, Li G, Ma J, Ma H (2020) Accelerated discovery of potential organic dyes for dye-sensitized solar cells by interpretable machine learning models and virtual screening. Sol RRL 4(6):2000110
16. Sahu H, Yang F, Ye X, Ma J, Fang W, Ma H (2019) Designing promising molecules for organic solar cells via machine learning assisted virtual screening. J Mater Chem A 7(29):17480–17488
17. Su A, Zhang C, She Y-B, Yang Y-F (2022) Exploring deep learning for metalloporphyrins: databases, molecular representations, and model architectures. Catalysts 12(11):1485
18. Su A, Zhang X, Zhang C, Ding D, Yang Y-F, Wang K, She Y-B (2023) Deep transfer learning for predicting frontier orbital energies of organic materials using small data and its application to porphyrin photocatalysts. Phys Chem Chem Phys 25(15):10536–10549
19. Li X, Zhang S-Q, Xu L-C, Hong X (2020) Predicting regioselectivity in radical C–H functionalization of heterocycles through machine learning. Angew Chem Int Ed 59(32):13253–13259
20. Shen H-M, Ye H-L, Ni J-Y, Wang K-K, Zhou X-Y, She Y-B (2023) Oxidation of αCH bonds in alkyl aromatics with O_2 catalyzed by highly dispersed cobalt(II) coordinated in confined reaction channel of porphyrin-based POFs with simultaneously enhanced conversion and selectivity. Chem Eng Sci 270:118472
21. Xu L-C, Frey J, Hou X, Zhang S-Q, Li Y-Y, Oliveira JCA, Li S-W, Ackermann L, Hong X (2023) Enantioselectivity prediction of pallada-electrocatalysed C–H activation using transition state knowledge in machine learning. Nat Synth 2(4):321–330
22. Pesciullesi G, Schwaller P, Laino T, Reymond J-L (2020) Transfer learning enables the molecular transformer to predict regio- and stereoselective reactions on carbohydrates. Nat Commun 11(1):4874
23. King-Smith E (2024) Transfer learning for a foundational chemistry model. Chem Sci 15(14):5143–5151
24. Zhuang F, Qi Z, Duan K, Xi D, Zhu Y, Zhu H, Xiong H, He Q (2020) A comprehensive survey on transfer learning. Proc IEEE 109(1):43–76
25. Devlin J, Chang M-W, Lee K, Toutanova K (2018) Bert: pre-training of deep bidirectional transformers for language understanding. arXiv preprint. arXiv:1810.04805
26. Schwaller P, Probst D, Vaucher AC, Nair VH, Kreutter D, Laino T, Reymond J-L (2021) Mapping the space of chemical reactions using attention-based neural networks. Nat Mach Intell 3(2):144–152
27. Yu J, Zhang C, Cheng Y, Yang Y-F, She Y-B, Liu F, Su W, Su A (2023) SolvBERT for solvation free energy and solubility prediction: a demonstration of an NLP model for predicting the properties of molecular complexes. Digit Discov 2(2):409–421
28. Zhang Y, Wang L, Wang X, Zhang C, Ge J, Tang J, Su A, Duan H (2021) Data augmentation and transfer learning strategies for reaction prediction in low chemical data regimes. Org Chem Front 8(7):1415–1423
29. Su A, Wang X, Wang L, Zhang C, Wu Y, Wu X, Zhao Q, Duan H (2022) Reproducing the invention of a named reaction: zero-shot prediction of unseen chemical reactions. Phys Chem Chem Phys 24(17):10280–10291
30. Schwaller P, Vaucher AC, Laino T, Reymond J-L (2021) Prediction of chemical reaction yields using deep learning. Mach Learn Sci Technol 2(1):015016
31. Mendez D, Gaulton A, Bento AP, Chambers J, De Veij M, Félix E, Magariños MP, Mosquera JF, Mutowo P, Nowotka M, Gordillo-Marañón M, Hunter F, Junco L, Mugumbate G, Rodriguez-Lopez M, Atkinson F, Bosc N, Radoux CJ, Segura-Cabrera A, Hersey A, Leach AR (2018) ChEMBL: towards direct deposition of bioassay data. Nucleic Acids Res 47(D1):D930–D940
32. Weininger D (1988) SMILES, a chemical language and information system. 1. Introduction to methodology and encoding rules. J Chem Inf Comput Sci 28(1):31–36
33. Hachmann J, Olivares-Amaya R, Atahan-Evrenk S, Amador-Bedolla C, Sánchez-Carrera RS, Gold-Parker A, Vogt L, Brockway AM, Aspuru-Guzik A (2011) The Harvard clean energy project: large-scale computational screening and design of organic photovoltaics on the world community grid. J Phys Chem Lett 2(17):2241–2251
34. Ørnsø KB, Garcia-Lastra JM, Thygesen KS (2013) Computational screening of functionalized zinc porphyrins for dye sensitized solar cells. Phys Chem Chem Phys 15(44):19478–19486
35. Ørnsø KB, Pedersen CS, Garcia-Lastra JM, Thygesen KS (2014) Optimizing porphyrins for dye sensitized solar cells using large-scale ab initio calculations. Phys Chem Chem Phys 16(30):16246–16254
36. Li Z, Omidvar N, Chin WS, Robb E, Morris A, Achenie L, Xin H (2018) Machine-learning energy gaps of porphyrins with molecular graph representations. J Phys Chem A 122(18):4571–4578
37. St. John PC, Phillips C, Kemper TW, Wilson AN, Guan Y, Crowley MF, Nimlos MR, Larsen RE (2019) Message-passing neural networks for high-throughput polymer screening. J Chem Phys 150(23):234111
38. Hao Z, Lu C, Huang Z, Wang H, Hu Z, Liu Q, Chen E, Lee C (2020) ASGN: an active semi-supervised graph neural network for molecular property prediction. In: Proceedings of the 26th ACM SIGKDD international conference on knowledge discovery & data mining, association for computing machinery: virtual event, CA, USA, pp 731–752
39. Joung JF, Han M, Jeong M, Park S (2020) Experimental database of optical properties of organic compounds. Sci Data 7(1):295
40. Joung JF, Han M, Hwang J, Jeong M, Choi DH, Park S (2021) Deep learning optical spectroscopy based on experimental database: potential applications to molecular design. JACS Au 1(4):427–438
41. Yu Z, Li Q, Ma Q, Ye W, An Z, Ma H (2023) Excited-state descriptors for high-throughput screening of efficient electro-fluorescent materials. Chem Mater 35(4):1827–1833
42. Capecchi A, Probst D, Reymond J-L (2020) One molecular fingerprint to rule them all: drugs, biomolecules, and the metabolome. J Cheminform 12(1):43
43. Probst D, Manica M, Nana Teukam YG, Castrogiovanni A, Paratore F, Laino T (2022) Biocatalysed synthesis planning using data-driven learning. Nat Commun 13(1):964
44. Shen WX, Zeng X, Zhu F, Wang YL, Qin C, Tan Y, Jiang YY, Chen YZ (2021) Out-of-the-box deep learning prediction of pharmaceutical properties by broadly learned knowledge-based molecular representations. Nat Mach Intell 3(4):334–343
45. Liu Y, Ott M, Goyal N, Du J, Joshi M, Chen D, Levy O, Lewis M, Zettlemoyer L, Stoyanov V (2019) Roberta: a robustly optimized bert pretraining approach. arXiv preprint. arXiv:1907.11692
46. Chithrananda S, Grand G, Ramsundar B (2020) ChemBERTa: large-scale self-supervised pretraining for molecular property prediction. arXiv preprint. arXiv:2010.09885
47. Probst D, Reymond J-L (2020) Visualization of very large high-dimensional data sets as minimum spanning trees. J Cheminform 12(1):1–13
48. Probst D, Reymond J-L (2018) FUn: a framework for interactive visualizations of large, high-dimensional datasets on the web. Bioinformatics 34(8):1433–1435
49. Ahmad W, Simon E, Chithrananda S, Grand G, Ramsundar B (2022) Chemberta-2: towards chemical foundation models. arXiv preprint. arXiv:2209.01712

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.