UC Irvine

UC Irvine Previously Published Works

Title

Improving molecular property prediction through a task similarity enhanced transfer learning strategy.

Permalink

https://escholarship.org/uc/item/20m5317q

Journal

iScience, 25(10)

Authors

Li, Han

Zhao, Xinyi

Li, Shuya

et al.

Publication Date

2022-10-21

DOI

10.1016/j.isci.2022.105231

Peer reviewed

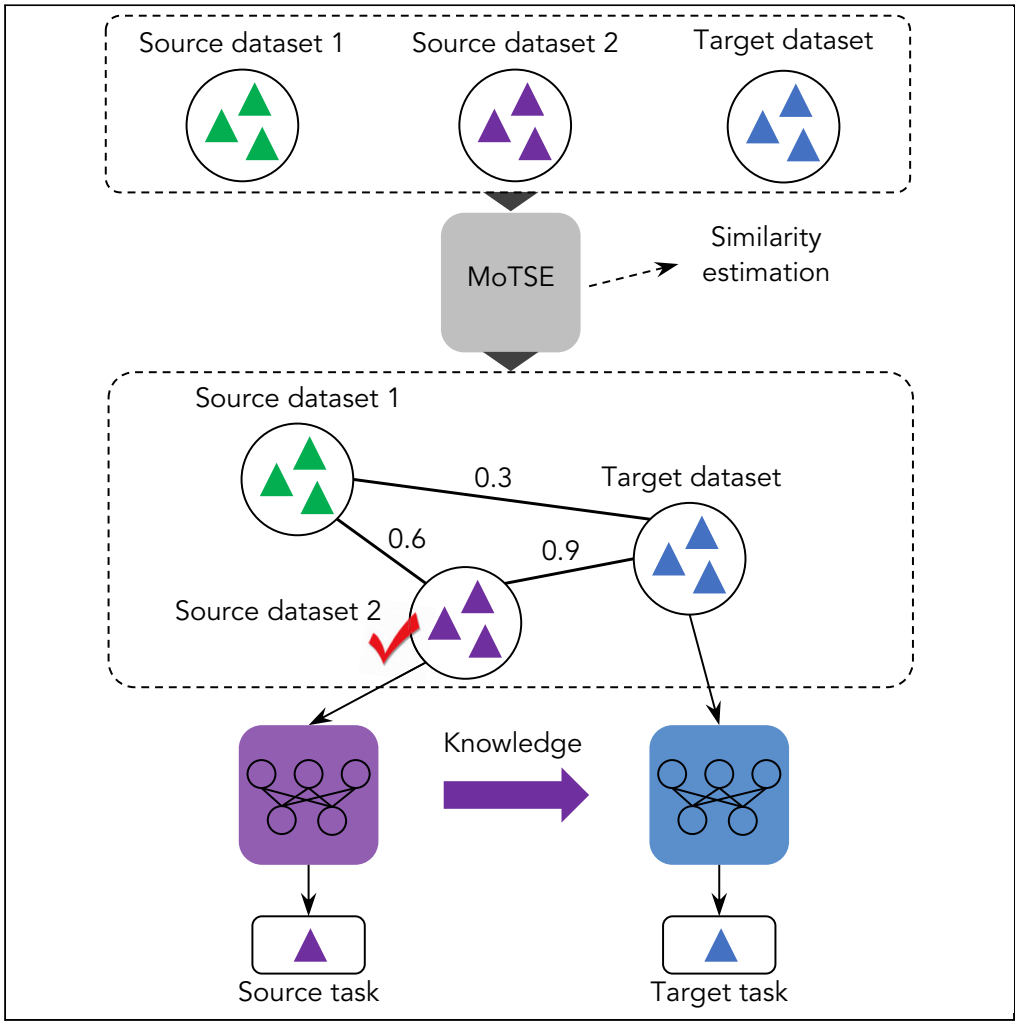
eScholarship.org Powered by the California Digital Library
University of California

iScience

CellPress
OPEN ACCESS

Article

Improving molecular property prediction through a task similarity enhanced transfer learning strategy



Han Li, Xinyi Zhao, Shuya Li, Fangping Wan, Dan Zhao, Jianyang Zeng

zhaodan2018@tsinghua.edu.cn (D.Z.)
zengjy321@tsinghua.edu.cn (J.Z.)

Highlights

MoTSE accurately measures similarity between molecular property prediction tasks

A novel transfer learning strategy to accurately predict molecular properties

An interpretable method to help understand relations between molecular properties

Li et al., iScience 25, 105231
October 21, 2022 © 2022 The Author(s).
https://doi.org/10.1016/j.isci.2022.105231

iScience

CellPress
OPEN ACCESS

Article

Improving molecular property prediction through a task similarity enhanced transfer learning strategy

Han Li,[1] Xinyi Zhao,[1] Shuya Li,[1] Fangping Wan,[2] Dan Zhao,[1,3,*] and Jianyang Zeng[1,3,*]

SUMMARY

Deeply understanding the properties (e.g., chemical or biological characteristics) of small molecules plays an essential role in drug development. A large number of molecular property datasets have been rapidly accumulated in recent years. However, most of these datasets contain only a limited amount of data, which hinders deep learning methods from making accurate predictions of the corresponding molecular properties. In this work, we propose a transfer learning strategy to alleviate such a data scarcity problem by exploiting the similarity between molecular property prediction tasks. We introduce an effective and interpretable computational framework, named MoTSE (Molecular Tasks Similarity Estimator), to provide an accurate estimation of task similarity. Comprehensive tests demonstrated that the task similarity derived from MoTSE can serve as useful guidance to improve the prediction performance of transfer learning on molecular properties. We also showed that MoTSE can capture the intrinsic relationships between molecular properties and provide meaningful interpretability for the derived similarity.

INTRODUCTION

With the development of high-throughput experimental techniques in the fields of biology and chemistry (Macarron et al., 2011), the number of available datasets of diverse molecular properties has increased significantly over the past few years (Ramakrishnan et al., 2014; Papadatos et al., 2015; Kim et al., 2016). This offers an unprecedented opportunity to design accurate computational models for molecular property prediction, thus facilitating the comprehension of molecular properties and accelerating the drug discovery process. However, as huge experimental efforts are often required for obtaining large-scale molecular property labels, the available data of the majority of the properties are still extremely scarce. For example, although the preprocessed ChEMBL dataset (Gaulton et al., 2012; Mayr et al., 2018) contains 1,310 bioassays and covers over 400K small molecules, the numbers of available labels of over 90% of the bioassays are below 1K. This data scarcity problem has limited the applications of data-driven computational models, especially deep learning models, in making accurate predictions of the corresponding molecular properties.

To alleviate the data scarcity problem, transfer learning strategies have been widely applied to improve the prediction performance of tasks with limited data in the field of computer vision (Zamir et al., 2018; Li et al., 2020; Chen and He, 2021). The general idea of transfer learning strategies is to transfer the knowledge learned from a source task with sufficient data to enhance the learning of a target task with limited data. The superior performance of transfer learning has also been well validated in molecular property prediction tasks (Simoes et al., 2018; Shen and Nicolaou, 2020; Cai et al., 2020; Li and Fourches, 2020). Nevertheless, the success of transfer learning is not always guaranteed. A number of studies have indicated that transfer learning can harm prediction performance (termed negative transfer) (Rosenstein et al., 2005; Fang et al., 2015; Wang et al., 2019b; Zhuang et al., 2021). It has been observed that negative transfer usually occurs when there exists only weak (or even no) similarity between the source and target tasks (Zhang et al., 2020). Therefore, to facilitate the effective applications of transfer learning in molecular property prediction and avoid the negative transfer problem, it is necessary to accurately measure the similarity between different molecular property prediction tasks.

It is generally hard to explicitly and manually measure the similarity between molecular property prediction tasks, even for experienced experts, as fully understanding the behaviors of molecules in the chemical and biological systems is extremely difficult owing to the high complexity of these systems. Fortunately,

[1]Institute for Interdisciplinary Information Sciences, Tsinghua University, Beijing 100084, China

[2]Machine Biology Group, Departments of Psychiatry and Microbiology, Institute for Biomedical Informatics, Institute for Translational Medicine and Therapeutics, Perelman School of Medicine, Departments of Bioengineering and Chemical and Biomolecular Engineering, School of Engineering and Applied Science, Penn Institute for Computational Science, University of Pennsylvania, Philadelphia, PA 19104, USA

[3]Lead contact

*Correspondence: zhaodan2018@tsinghua.edu.cn (D.Z.), zengjy321@tsinghua.edu.cn (J.Z.)

https://doi.org/10.1016/j.isci.2022.105231

This is an open access article under the CC BY license (http://creativecommons.org/licenses/by/4.0/).

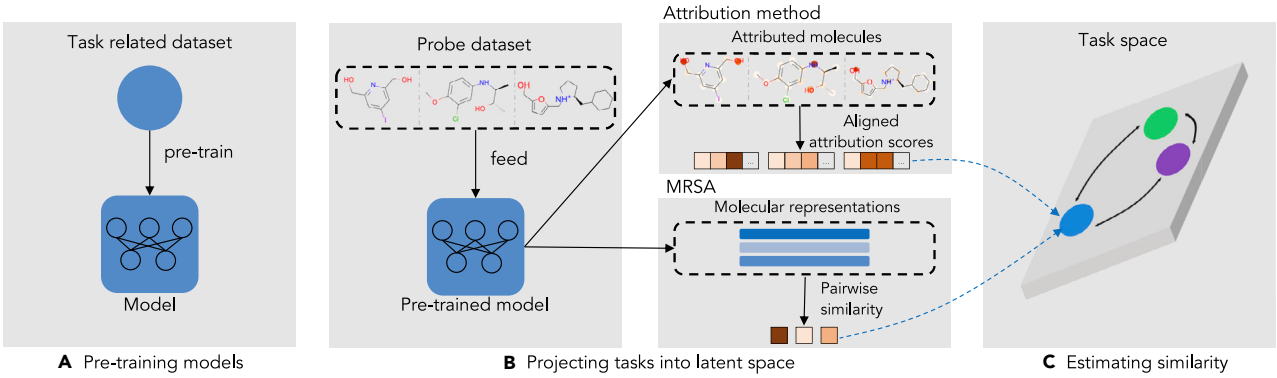


Figure 1. An illustrative diagram of MoTSE

(A) Given a task, MoTSE first pre-trains a GNN model using the corresponding dataset in a supervised manner.

(B) By means of a probe dataset, MoTSE extracts the task-related knowledge from the pre-trained GNN and projects the task into a latent task space. The knowledge extraction is achieved by two methods: an attribution method extracting the task-related local knowledge by assigning importance scores to atoms in molecules; and a molecular representation similarity analysis (MRSA) method extracting the task-related global knowledge by pair-wisely measuring the similarity between molecular representations.

(C) Finally, MoTSE calculates the similarity between tasks by measuring the distances between the corresponding vectors in the task space.

data-driven computational methods can provide an implicit way to enable us to define and measure task similarity. The seminal work of Taskonomy (Zamir et al., 2018) has made a pioneering attempt toward modeling the similarity between computer vision tasks through a deep learning approach. The results have shown that incorporating the similarity derived from Taskonomy can improve the performance of transfer learning on computer vision tasks. In addition, the similarity tree constructed according to the derived similarity is highly consistent with human conceptions, indicating that such approaches can potentially capture the intrinsic relationships between tasks. This thus inspires us to develop a computational method for estimating the similarity between molecular property prediction tasks, which can not only guide the source task selection to avoid negative transfer in transfer learning but also provide useful hints in understanding the relationships between tasks.

To this end, we propose MoTSE, an interpretable computational framework, to efficiently measure the similarity between molecular property prediction tasks. MoTSE is based on the assumption that two tasks should be similar if the hidden knowledge learned by their task-specific models is close to each other. More specifically, MoTSE first pre-trains a graph neural network (GNN) model for each task. Then an attribution method and a molecular representation similarity analysis (MRSA) method are introduced to represent the hidden knowledge enclosed in the pre-trained GNNs as embedded vectors and project individual tasks into a unified latent space. Finally, MoTSE calculates the distances between the vectors in the latent space to derive the similarity between different tasks. Based on the task similarity derived from MoTSE, we design a novel transfer learning strategy to enhance the learning of the molecular property prediction tasks with limited data.

Our extensive computational tests demonstrated that the task similarity estimated by MoTSE can successfully guide the source task selection in transfer learning, with superior prediction performance over a number of baseline methods, including multitask learning, training from scratch, and nine state-of-the-art self-supervised learning methods, on several molecular property datasets from various domains. Meanwhile, by applying MoTSE to a dataset measuring the physical chemistry properties and a dataset measuring the bio-activities against cytochrome P450 isozymes, we also demonstrated that MoTSE was able to capture the intrinsic relationships between molecular properties and provide meaningful interpretability for the derived similarity.

RESULTS

Overall design of MoTSE

Figure 1 illustrates the overall architecture of MoTSE. Given a set of molecular property prediction tasks with the corresponding datasets, MoTSE estimates the task similarity via the following three main steps: (1) Representing molecules as graphs, where nodes represent atoms and edges represent covalent bonds

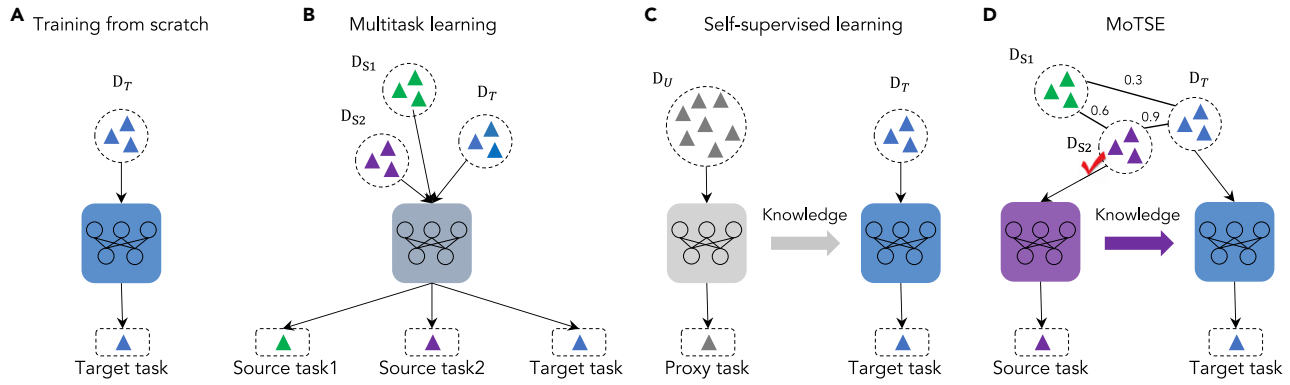


Figure 2. Schematic illustration of different learning strategies

(A) Training from scratch directly trains a model on the dataset of each target task without exploiting any extra knowledge. (B) Multitask learning learns the target task and source tasks simultaneously. (C) Self-supervised learning first leverages a proxy task to learn general knowledge from a large-scale unlabeled dataset and then finetunes the pre-trained model on the dataset of the target task. (D) MoTSE-guided transfer learning first pre-trains a model on the most similar task with the target task according to the task similarity estimated by MoTSE and then finetuned the pre-trained model on the dataset of the target task. D_T stands for the dataset for the target task, D_{S1} and D_{S2} stand for the datasets for the source tasks, and D_U stands for the large-scale unlabeled dataset. The numbers between the datasets represent the similarity estimated by MoTSE between the corresponding tasks.

(see Figure S1), MoTSE pre-trains a graph neural network (GNN) model on the dataset for each task in a supervised manner. (2) By means of a probe dataset (i.e., a set of unlabeled molecules, see STAR Methods for more details), MoTSE extracts the task-related knowledge from the pre-trained GNNs and then projects the tasks into a unified latent task space. The knowledge extraction is achieved by an attribution method and a molecular representation similarity analysis (MRSA) method. These two methods are effectively complementary to each other: the attribution method extracts the local knowledge by assigning importance scores to atoms in molecules and the MRSA method extracts the global knowledge by pair-wisely measuring the similarity between molecular representations. (3) MoTSE estimates the similarity between tasks by calculating the distances between the corresponding vectors in the projected latent task space.

Based on the task similarity derived from MoTSE, we design a novel transfer learning strategy to improve the prediction performance for molecular properties with limited data. More specifically, given a target task, we first select the most similar task according to the task similarity estimated by MoTSE as its source task and then finetune the model pre-trained on the source task to exploit its related knowledge to enhance the learning of the target task. As GNN models have shown superior capability in learning hidden knowledge and modeling various kinds of molecular properties (Gilmer et al., 2017; Li et al., 2019; Xiong et al., 2019), here we also adopt the GNN models to capture the hidden knowledge contained in individual tasks. Note that, MoTSE is orthogonal to different GNN architectures. We use graph convolutional networks (GCNs) (Kipf and Welling, 2016) in our computational experiments if not specially specified (see Figure S2 for an illustrative diagram for our model architecture). More details about MoTSE, the transfer learning strategy, the model architecture, and the training process can be found in STAR Methods.

The MoTSE-guided transfer learning strategy outperforms baseline methods

We systematically evaluated the performance of our MoTSE-guided transfer learning strategy on molecular property prediction. We made comparison with eleven baseline methods with different learning strategies, including multitask learning (MT), training from scratch (Scratch), and nine state-of-the-art self-supervised learning methods, i.e., EdgePred (Hamilton et al., 2017), DGI (Velickovic et al., 2019), Masking (Hu et al., 2020), ContextPred (Hu et al., 2020), JOAO (You et al., 2021), EdgePred$_{sup}$ (Hu et al., 2020), Masking$_{sup}$ (Hu et al., 2020), ContextPred$_{sup}$ (Hu et al., 2020) and DGI$_{sup}$ (Hu et al., 2020) (see STAR Methods for more details about these baseline methods). A schematic illustration of our MoTSE-guided transfer learning strategy and other learning schemes is shown in Figure 2.

We first applied the following two representative datasets QM9 (Ramakrishnan et al., 2014) and PCBA (Ramsundar et al., 2015) for performance evaluation, in which the QM9 dataset measured the quantum chemical properties and the PCBA dataset measured the bio-activities of small molecules (see Table S1

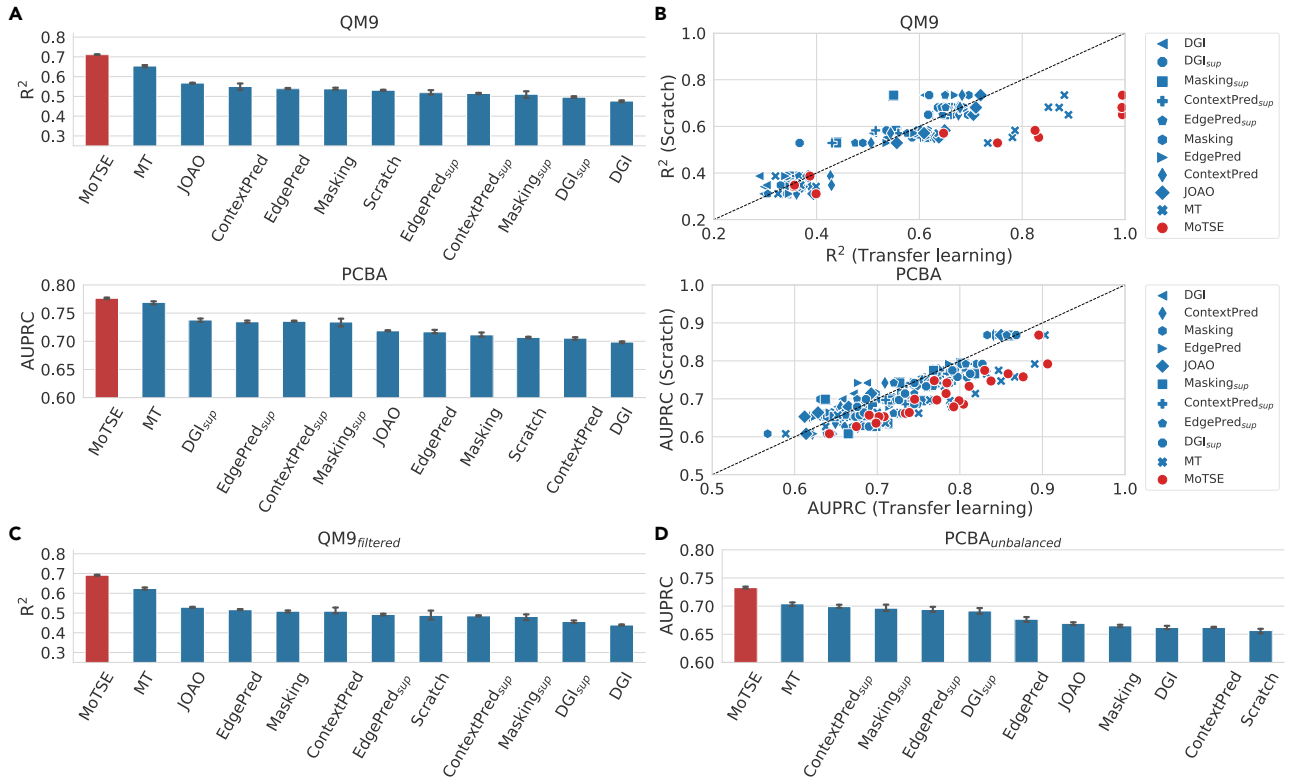


Figure 3. MoTSE outperforms baseline methods and alleviates negative transfer on the QM9 and PCBA datasets

(A) The prediction performance of MoTSE and eleven baseline methods on the QM9 and PCBA datasets, measured in terms of R^2 and AUPRC, respectively.

(B) The prediction performance of eleven transfer learning methods versus that of the Scratch method on the QM9 and PCBA datasets.

(C) The comparison results of R^2 between MoTSE and eleven baseline methods on the QM9 dataset after filtering every molecule from the test set if it has a Tanimoto similarity score greater than 0.8 to any molecule in the training set (also see Figure S4A).

(D) The comparison results of AUPRC between MoTSE and eleven baseline methods on an unbalanced PCBA dataset with only 10% positive samples (also see Figure S4B).

and STAR Methods for more details about the datasets used in our tests). To evaluate the effectiveness of different learning strategies, we further preprocessed the datasets to (1) mimic a specific scenario of transfer learning, in which the data size of the source task was relatively larger than that of the target task, and (2) reduce the influence of other factors (e.g., data size) that might affect the performance of transfer learning and thus only focus on the effect of learning strategies themselves. In particular, we first created two subsets $QM9_{10k}$ and $PCBA_{10k}$ as the datasets for the source tasks, in which each task had about 10,000 data samples, and then randomly partitioned the datasets into training, validation, and test sets with a ratio of 8:1:1. Next, we constructed another two subsets $QM9_{1k}$ and $PCBA_{1k}$ as the datasets for the target tasks by: (1) constructing training and validation sets by sampling 800 and 100 data samples from the corresponding training and validation sets of $QM9_{10k}$ and $PCBA_{10k}$, respectively, to avoid data leakage in the transfer learning; and (2) sharing the test sets with $QM9_{10k}$ and $PCBA_{10k}$, respectively, for an accurate performance evaluation (see Figure S3 for an illustrative diagram of the dataset generation process).

For each dataset of QM9 and PCBA, we sequentially treated one task in the dataset as a target task and the others as the source tasks. MoTSE measured the task similarity based on the models trained on the $QM9_{1k}$ and $PCBA_{1k}$ datasets. We performed three repeated tests with different random seeds and reported the averaged R^2 and AUPRC scores on the QM9 and PCBA datasets, respectively (see Figure 3A). We found that MoTSE can make accurate predictions and outperformed all the baseline methods. As mentioned previously, transfer learning can lead to negative transfer (i.e., the performance of a transfer learning method is worse than that of training from scratch) when the source task is not properly defined. We plotted the prediction performance of each task from eleven transfer learning methods versus that from the training from scratch method (see Figure 3B). We observed that MoTSE perfectly avoided the negative transfer problem

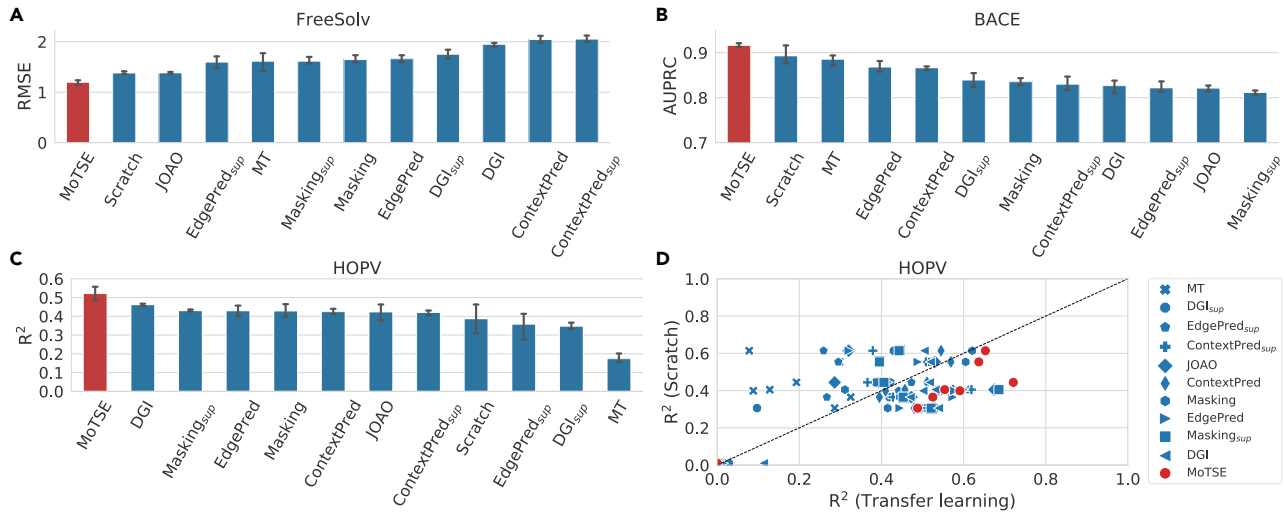


Figure 4. The prediction performance of MoTSE and baseline methods on the FreeSolv, BACE, and HOPV datasets

(A) The comparison results between MoTSE and eleven baseline methods on the FreeSolv dataset, measured in terms of root-mean-square-error (RMSE).

(B) The comparison results between MoTSE and eleven baseline methods on the BACE dataset, measured in terms of AUPRC.

(C) The comparison results between MoTSE and eleven baseline methods on the HOPV dataset, measured in terms of R^2.

(D) The prediction performance of eleven transfer learning methods versus that of the Scratch method on the HOPV dataset, measured in terms of R^2.

on the QM9 and PCBA datasets, while all the baseline methods suffered from this problem to varying degrees.

Next, we benchmarked MoTSE in more challenging scenarios. For the QM9 dataset, we filtered the test set of $QM9_{1k}$ by excluding every molecule from the test set if it had a Tanimoto similarity score greater than 0.8 to any molecule in the training set (denoted by $QM9_{filtered}$). For the PCBA dataset, we generated an unbalanced dataset with only 10% positive samples (denoted by $PCBA_{unbalanced}$). We found that MoTSE still consistently outperformed baseline methods (see Figures 3C and 3D), and overcame negative transfer on all the tasks of these two representative challenging test cases (see Figure S4).

We also evaluated our method in more practical scenarios in which the source tasks and the target tasks were from different domains. More specifically, we first employed the FreeSolv dataset (Mobley and Guthrie, 2014), which produced a regression task measuring the solubility of 614 molecules. We derived the task similarity using the $QM9_{1k}$ and FreeSolv datasets and used the tasks from the $QM9_{10k}$ dataset as the source tasks. We employed MoTSE to enhance the transfer learning process and made a comparison with baseline methods. As shown in Figure 4A, MoTSE achieved better performance in comparison with baseline methods. Then we tested MoTSE on the BACE dataset (Subramanian et al., 2016), which measured whether each of 1513 molecules can act as an inhibitor of human β-secretase 1 (BACE-1). We first derived the task similarity using the $PCBA_{1k}$ and BACE datasets, and then used the tasks from the $PCBA_{10k}$ dataset as the source tasks. The comparison results between MoTSE and the baseline methods are shown in Figure 4B, which showed that our method still outperformed baseline methods. These results indicated that MoTSE can still accurately model the underlying similarity between molecular property prediction tasks even for the properties from different domains.

To further evaluate the ability of MoTSE in enhancing the prediction of molecular properties on extremely small datasets, we also accessed its performance on the HOPV dataset (Lopez et al., 2016), which contained only 350 molecules and measured eight quantum chemical properties. Here, we employed the tasks in the $QM9_{10k}$ dataset as the source ones and used MoTSE to select the source task for each target task in the HOPV dataset. In comparison with baseline methods, MoTSE made more accurate predictions on the HOPV dataset and also achieved better results in addressing the negative transfer problem (see Figures 4C and 4D).

We also conducted additional tests to investigate the impact of the sizes of target and source datasets on the prediction performance of MoTSE (see Figures S6). Our analyses showed that the prediction

performance of MoTSE was improved with the increase of the sizes of source and target datasets and MoTSE consistently outperformed Scratch, which demonstrated the robustness of MoTSE to the sizes of source and target datasets. Note that, MoTSE can still offer performance gain even when the source datasets only contain equal or fewer data samples than the target dataset (see Figures S6C and S6D). Based on these observations, we empirically recommended applying MoTSE on those target datasets with relatively limited data samples (e.g., less than 3,000) and employing source datasets that contain more data samples than target datasets, as MoTSE can achieve relatively larger performance gain under these conditions. Furthermore, we sought to define a proper threshold value of the similarity between the source task and target task that can effectively enable MoTSE to guide the transfer learning process. We first plotted the similarity between source and the target tasks versus the performance improvement on the QM9 and PCBA datasets, respectively. As shown in Figure S7, MoTSE achieved better prediction performance when the similarity between the source and target tasks was larger than 0.7.

The task similarity estimated by MoTSE is generalizable across models with different architectures and datasets with different distributions

We next sought to explore whether the similarity estimated by MoTSE was generalizable across models with different architectures and datasets with different distributions, that is, whether the task similarity derived from MoTSE equipped with a certain model or on a certain dataset was generalizable to enhance the learning of other model architectures or datasets with different data distributions.

We first considered three models with different architectures in the tests, including a graph attention network (denoted as GAT) (Veličković et al., 2017), an ECFP (i.e., extended connectivity fingerprint) (Rogers and Hahn, 2010) based fully connected network (denoted as FCN) and a SMILES (i.e., simplified molecular input line entry specification) (Weininger, 1988) based recurrent neural network (denoted as RNN) (more details about these three types of models can be found in STAR Methods). Then, with the guidance of the similarity estimated by MoTSE equipped with the GCN model, we evaluated the transfer learning performance of the above three types of models on the QM9 and PCBA datasets and made comparisons with the baseline methods. Here we omitted the results of the nine self-supervised learning strategies on the FCN and RNN models, as they were particularly designed for GNNs and cannot be easily generalized to the FCN and RNN models. We observed that MoTSE consistently achieved significant improvement on the QM9 and PCBA datasets using different model architectures in comparison with all the baseline methods (see Figures 5A-5C). Moreover, we constructed similarity trees of tasks in the QM9 dataset using the hierarchical agglomerative clustering algorithm (Jain et al., 1999) according to the task similarity estimated based on GCN and GAT, respectively (see Figure S8). The similarity trees were highly consistent with each other. These results indicated that the task similarity estimated by MoTSE was generalizable across different model architectures.

Next, to evaluate the generalizability of the similarity estimated by MoTSE across datasets with different data distributions, we employed the Alchemy dataset (Chen et al., 2019), which shared the same tasks but had a different data distribution compared with the QM9 dataset, that is, the QM9 dataset contained molecules comprising up to nine non-hydrogen atoms while the molecules in the Alchemy dataset consisted of nine to fourteen non-hydrogen atoms. We first preprocessed the Alchemy dataset and created Alchemy_{10k} and Alchemy_{1k} following the preprocessing process shown in Figure S3. Then, for each source task, we pre-trained the models on the Alchemy_{10k} dataset. Next, for each target task in the Alchemy_{1k} dataset, we selected the source task from the Alchemy_{10k} dataset according to the task similarity estimated based on the QM9_{1k} dataset and fine-tuned on the Alchemy_{1k} dataset. We found that MoTSE still outperformed the baseline methods in this case (see Figure 5D). Moreover, we constructed the similarity trees according to the similarity estimated by MoTSE on the QM9 and Alchemy datasets, respectively. We observed that the structures of the derived similarity trees were highly consistent with each other (see Figure 5E).

These results demonstrated that the task similarity derived from MoTSE was generalizable across models with different architectures and datasets with different data distributions, which indicated that MoTSE can capture the model and dataset independent similarity between molecular property prediction tasks. Therefore, once the similarity between molecular property prediction tasks was estimated by MoTSE, it can be directly applied to enhance the learning of diverse model architectures and novel datasets in future studies.

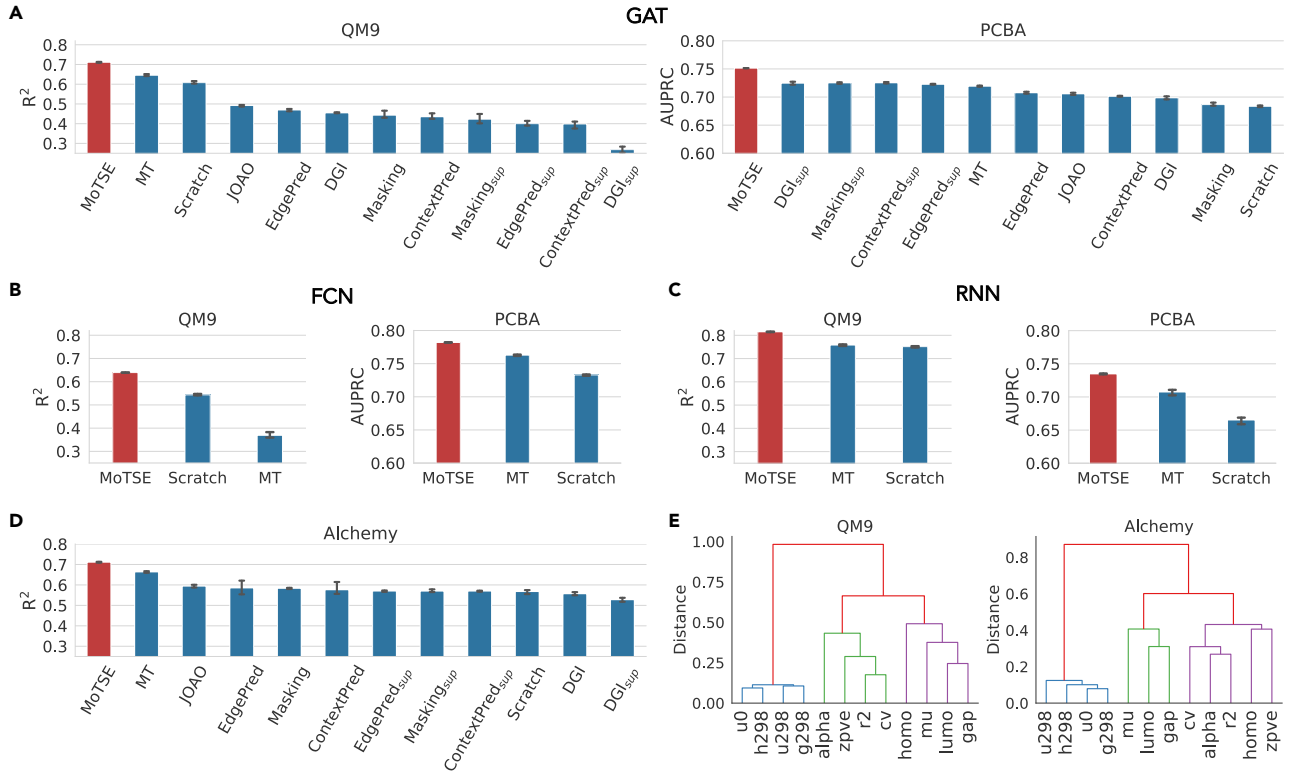


Figure 5. The task similarity derived from MoTSE is generalizable across models with different architectures and datasets with different data distributions

(A-C) The comparison results between MoTSE and baseline methods on the QM9 and PCBA datasets (measured in terms of R^2 and AUPRC, respectively), using the graph attention network (GAT), fully connected network (FCN), and recurrent neural network (RNN) models, respectively.

(D) The comparison results between MoTSE and baseline methods on the Alchemy dataset, measured in terms of R^2.

(E) The similarity trees constructed based on the task similarity estimated by MoTSE on the QM9 and Alchemy datasets, respectively.

Task similarity derived from MoTSE reflects intrinsic relationships between physical chemistry properties

Next, we asked whether the task similarity derived from MoTSE was consistent with the intrinsic relationships between molecular properties. We constructed a dataset containing 10K molecules labeled with four well-studied physical chemistry tasks, including NHA (number of hydrogen acceptors contained in a molecule), NHD (number of hydrogen donors contained in a molecule), NOcount (number of nitrogen (N) and oxygen (O) atoms contained in a molecule), and NHOHCount (number of N and O atoms that are covalently bonded with hydrogens in a molecule) (see STAR Methods for more details of this dataset). Then we applied MoTSE to estimate the similarity between these tasks.

From the chemical perspective, NHD is expected to be more similar to NHOHCount than NOCount, as only those N and O atoms with covalently bonded hydrogens can serve as hydrogen donors. NHA is expected to be more similar to NOCount than NHOHCount, as those N and O atoms both with or without covalently bonded hydrogens can be hydrogen acceptors. We observed that the task similarity derived from MoTSE was entirely consistent with these facts (see Figure 6).

Meanwhile, we visualized the importance scores of the atoms derived from the attribution method employed in MoTSE (see Figure 6). We found that MoTSE precisely assigned high importance scores to those target atoms related to the properties. For example, the N and O atoms were emphasized for the NOCount task, and the NH and OH atoms with hydrogen bonds were emphasized for the NHOHCount task. We also found that similar tasks tended to assign similar importance scores to the same atoms in molecules. For instance, NHD and NHOHCount both assigned higher importance scores to the N and O atoms with covalently bonded hydrogens. These observations interpreted how MoTSE estimated similarity between tasks

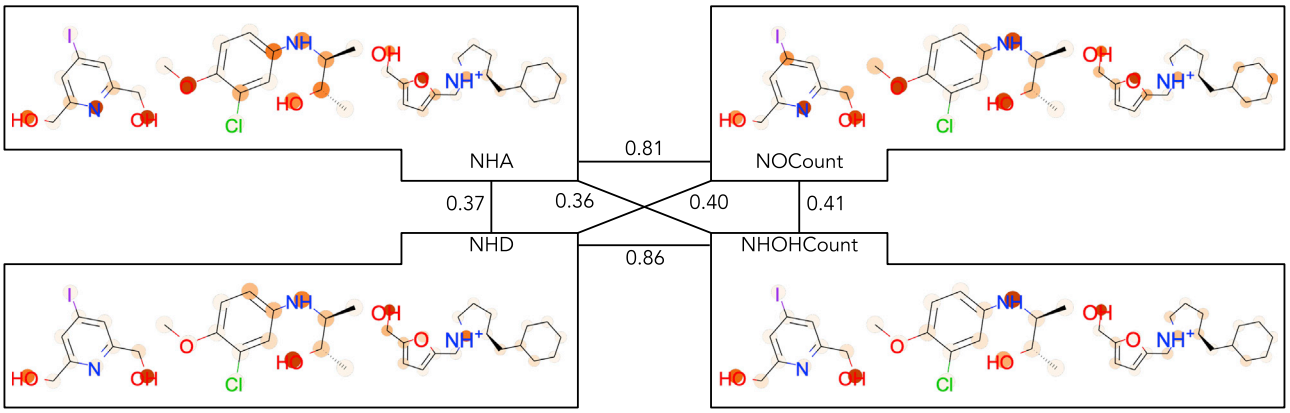


Figure 6. The similarity estimated by MoTSE between four physical chemistry tasks and the example molecules with importance scores assigned by the attribution method employed in MoTSE

The numbers between tasks denote the task similarity derived from MoTSE. In the visualized molecules, darker colors represent higher importance scores. See the main text for the definitions of the four physical chemistry task.

and indicated that our method was able to capture the intrinsic similarity between tasks by exploiting the chemical concepts behind the corresponding molecular properties.

Measuring and interpreting similarity between the tasks of estimating the bio-activities of molecules against cytochrome P450 isozymes

To further evaluate the ability of MoTSE in estimating and interpreting the similarity between molecular properties, we carried out a more challenging experiment, which included five tasks of predicting the bio-activities of small molecules against cytochrome P450 isozymes. The cytochrome P450 (CYP) family plays important roles in drug metabolism, especially for five isozymes—1A2, 2C9, 2C19, 2D6 and 3A4 (Williams et al., 2004; De Montellano, 2005). Here, we obtained the binary bio-activity labels between around 17K molecules and the above five CYP isozymes from the preprocessed ChEMBL dataset (Mayr et al., 2018). We then applied MoTSE to estimate the similarity of the tasks.

According to the similarity estimated by MoTSE, we first constructed a similarity tree using the hierarchical agglomerative clustering algorithm (Jain et al., 1999) (see Figure 7A). We observed that the similarity between CYP2C9 and CYP2C19 estimated by MoTSE was the highest among all pairs of CYP isozymes, which was consistent with the fact that CYP 2C9 and 2C19 genetically shared the most (91%) sequence homology (Attia et al., 2014). Meanwhile, we found that the structure of this tree was exactly the same as that derived by self-organizing maps (SOMs) (Schneider and Schneider, 2003; Selzer and Ertl, 2006) offered in previous research (Veith et al., 2009), in which the SOMs of individual isozymes were constructed based on the structural similarity of molecules and reflected the activity patterns (i.e., the scaffolds enriched in active or inactive molecules) of corresponding isozymes. According to this observation, we expected that MoTSE may capture the activity patterns of the CYP bioactivity prediction tasks and fully exploit such knowledge to estimate the similarity between these tasks.

To validate this hypothesis, we further visualized several molecules with the importance scores assigned by the attribution method employed in MoTSE. As shown in Figures 7B-7F, we found that similar tasks tended to share the same active patterns. For example, CYP 2C9 and CYP 2C19 shared the same five active patterns. In addition, the active patterns highlighted by our attribution method can be supported by previous research (Kho et al., 2006; Veith et al., 2009; Lee et al., 2017). For example, the substructure in Figure 7B was also previously considered as an active pattern of CYP 2C9, CYP 2C19 and CYP 3A4 by substructure searching (Veith et al., 2009) and fingerprint analysis (Lee et al., 2017).

The above results demonstrated that MoTSE can successfully extract task-related knowledge and thus accurately estimate the intrinsic similarity between the tasks (e.g., the similarity between active patterns and the genetic similarity between CYP isozymes). Therefore, MoTSE can potentially provide a novel perspective to help understand the mechanisms behind the bio-activities of small molecules.

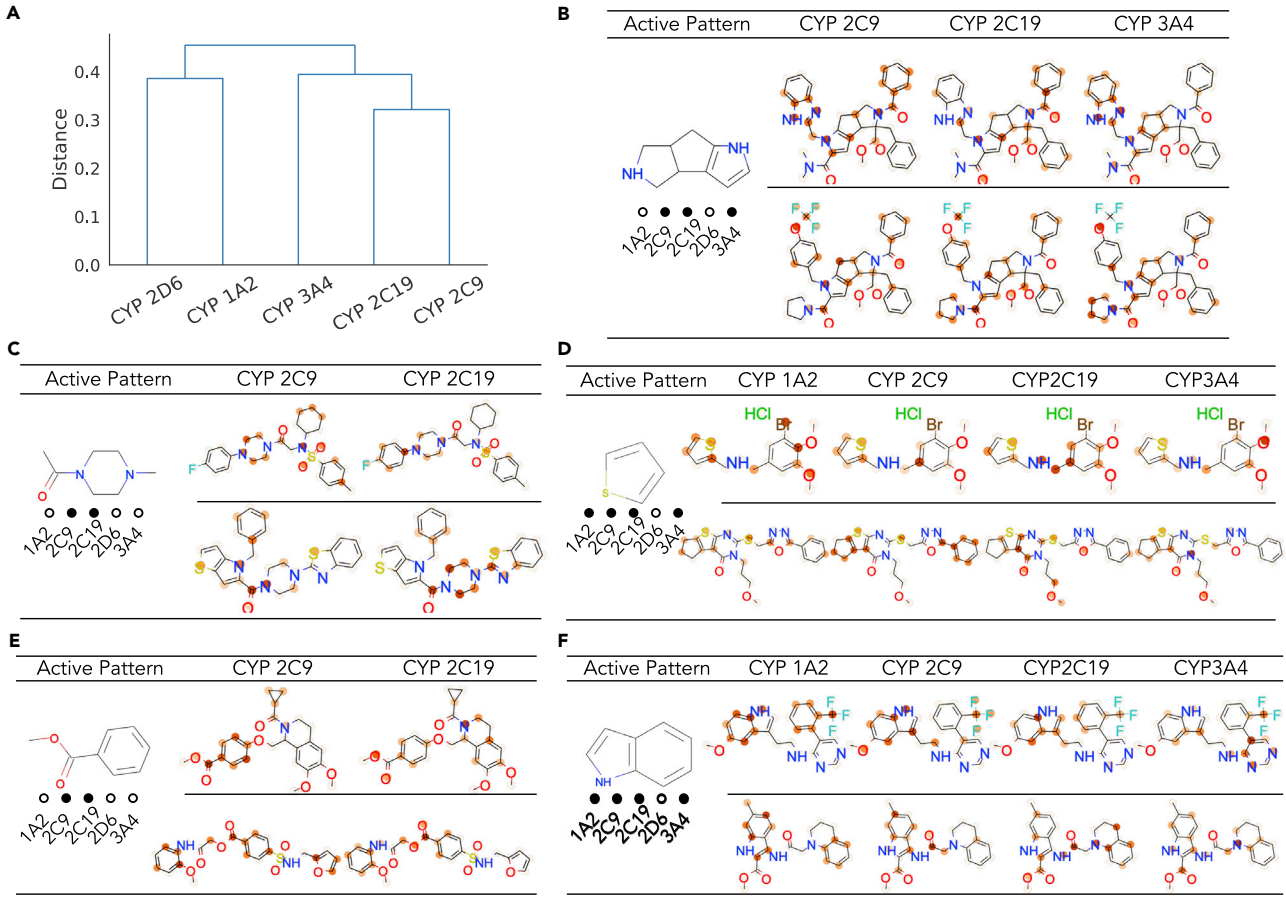


Figure 7. Measuring and interpreting the similarity between the tasks of estimating the bio-activities of molecules against cytochrome P450 isozymes

(A) The task similarity tree constructed using the similarity estimated by MoTSE.

(B-F) Five active patterns highlighted by our attribution method. The filled or hollow circles below a functional group represent whether the corresponding functional group is an active pattern for individual isozymes 1A2, 2C9, 2C19, 2D6 and 3A4 or not. The functional groups are shown on the left, while the molecules with the active patterns are shown on the right.

Conclusion

In this article, we present MoTSE, a computational method to efficiently estimate the similarity between molecular property prediction tasks. Specifically, we first pre-train a GNN to automatically capture task-related knowledge from the corresponding datasets. Then we employ the attribution method and the MRSA method to, respectively, extract both local and global knowledge contained in the pre-trained GNNs with the help of a probe dataset and project individual tasks as vectors into a unified latent task space. Finally, the similarity between the tasks can be measured by calculating the distances between the corresponding embedded vectors in the latent task space. The derived task similarity can be applied to design an accurate transfer learning strategy to enhance the prediction of molecular properties with limited data sizes. To ensure effective transfer learning, we empirically recommend applying MoTSE on the target datasets with limited data samples (e.g., less than 3,000) and employing the source datasets that contain more data samples than the target datasets. We also recommend selecting tasks with a similarity greater than 0.7 to the target task as the source tasks. In comparison with current transfer learning strategies, which attempt to leverage one proxy task to learn knowledge that can be generalized to molecular properties from different domains (i.e., self-supervised learning) or arbitrarily learn the target task and multiple source tasks simultaneously (i.e., multitask learning), our proposed transfer learning strategy offers a more reasonable and effective way to select a proper source task for each target task individually, thus fully taking advantage of the knowledge from the source task with sufficient data samples.

Comprehensive test results showed that the MoTSE-guided transfer learning strategy significantly outperformed the baseline learning strategies in predicting molecular properties and avoiding the negative transfer problem, especially on those datasets with limited data. MoTSE was also robust to different sizes of target and source datasets. Moreover, we validated that MoTSE achieved superior performance in the scenarios where the source and target tasks were from different domains. All these results demonstrated that MoTSE can be applied to molecular property prediction tasks from various scenarios. Therefore, MoTSE can provide a useful tool to fully exploit the increasing number of large-scale molecular property datasets to enhance the learning of properties with only limited training data, which is of great importance to accelerate the early stage of finding drug candidate molecules. In addition, we demonstrated that MoTSE can capture the intrinsic relationships between molecular properties and provide meaningful interpretability for the derived similarity, which can potentially help biologists/chemists understand the underlying mechanisms behind molecular properties.

Limitations of the study

In our proposed learning strategy, we select the most similar source task to enhance the learning of one target task in a one-to-one transfer manner (i.e., transferring one source task to one target task). Although the test results have demonstrated the superior performance of such a strategy, there is still room for further explorations about improving the learning strategy. For example, we can design effective strategies to simultaneously take advantage of the top-n ($n > 1$) similar tasks in the pre-training stage. In addition, a curriculum learning strategy can be designed by building effective learning paths (e.g., source task A → source task B → target task) based on the similarity derived from MoTSE. These points were not fully explored in our current work but will be interesting directions in future studies.

STAR★METHODS

Detailed methods are provided in the online version of this paper and include the following:

- KEY RESOURCES TABLE
- RESOURCE AVAILABILITY
  - Lead contact
  - Materials availability
  - Data and code availability
- METHOD DETAILS
  - Notation and problem setting
  - Key steps of MoTSE
  - Implementation of MoTSE
  - Datasets and data processing
  - Learning strategies
  - Model configurations

SUPPLEMENTAL INFORMATION

Supplemental information can be found online at https://doi.org/10.1016/j.isci.2022.105231.

ACKNOWLEDGMENTS

This work was supported in part by the National Key Research and Development Program of China (2021YFF1201300), the National Natural Science Foundation of China (61872216, T2125007 to JZ, 31900862 to DZ), the Turing AI Institute of Nanjing, and the Tsinghua-Toyota Joint Research Fund.

AUTHOR CONTRIBUTIONS

Conceptualization, H.L., D.Z., and J.Z.; Methodology, H.L. and X.Z.; Investigation, H.L., X.Z., S.L., F.W., D.Z., and J.Z.; Writing - Original Draft, H.L., X.Z., and J.Z.; Writing - Review & Editing, H.L., X.Z., S.L., F.W., D.Z., and J.Z.; Funding Acquisition, D.Z. and J.Z.; Resources, J.Z.; Supervision, H.L., D.Z., and J.Z.

DECLARATION OF INTERESTS

J.Z. is a founder of the Silexon AI Technology Co. Ltd and has an equity interest.

Received: June 5, 2022
Revised: September 2, 2022
Accepted: September 23, 2022
Published: October 21, 2022

REFERENCES

Agarap, A.F. (2018). Deep learning using rectified linear units (relu). Preprint at arXiv. https://doi.org/10.48550/arXiv.1803.08375.

Arús-Pous, J., Johansson, S.V., Prykhodko, O., Bjerrum, E.J., Tyrchan, C., Reymond, J.L., Chen, H., and Engkvist, O. (2019). Randomized smiles strings improve the quality of molecular generative models. J. Cheminformatics *11*, 1–13. https://doi.org/10.1186/s13321-019-0393-0.

Attia, T.Z., Yamashita, T., Hammad, M.A., Hayasaki, A., Sato, T., Miyamoto, M., Yasuhara, Y., Nakamura, T., Kagawa, Y., Tsujino, H., et al. (2014). Effect of cytochrome p450 2c19 and 2c9 amino acid residues 72 and 241 on metabolism of tricyclic antidepressant drugs. Chem. Pharm. Bull. *62*, 176–181. https://doi.org/10.1248/cpb.c13-00800.

Cai, C., Wang, S., Xu, Y., Zhang, W., Tang, K., Ouyang, Q., Lai, L., and Pei, J. (2020). Transfer learning for drug discovery. J. Med. Chem. *63*, 8683–8694. https://doi.org/10.1021/acs.jmedchem.9b02147.

Chen, G., Chen, P., Hsieh, C.Y., Lee, C.K., Liao, B., Liao, R., Liu, W., Qiu, J., Sun, Q., Tang, J., et al. (2019). Alchemy: a quantum chemistry dataset for benchmarking ai models. Preprint at arXiv. https://doi.org/10.48550/arXiv.1906.09427.

Chen, X., and He, K. (2021). Exploring simple siamese representation learning. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 15750–15758. https://doi.org/10.1109/CVPR46437.2021.01549.

De Montellano, P.R.O. (2005). Cytochrome P450: Structure, Mechanism, and Biochemistry (Springer Science & Business Media). https://doi.org/10.1021/ja041050x.

Dwivedi, K., and Roig, G. (2019). Representation similarity analysis for efficient task taxonomy & transfer learning. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 12387–12396. https://doi.org/10.1109/CVPR.2019.01267.

Fang, M., Guo, Y., Zhang, X., and Li, X. (2015). Multi-source transfer learning based on label shared subspace. Pattern Recognit. Lett. *51*, 101–106. https://doi.org/10.1016/j.patrec.2014.08.011.

Gaulton, A., Bellis, L.J., Bento, A.P., Chambers, J., Davies, M., Hersey, A., Light, Y., McGlinchey, S., Michalovich, D., Al-Lazikani, B., and Overington, J.P. (2012). Chembl: a large-scale bioactivity database for drug discovery. Nucleic Acids Res. *40*, D1100–D1107. https://doi.org/10.1093/nar/gkr777.

Gilmer, J., Schoenholz, S.S., Riley, P.F., Vinyals, O., and Dahl, G.E. (2017). Neural message passing for quantum chemistry. Preprint at arXiv. https://doi.org/10.5555/3305381.3305512.

Goh, G.B., Hodas, N., Siegel, C., and Vishnu, A. (2018). Smiles2vec: predicting chemical properties from text representations. Preprint at arXiv. https://doi.org/10.48550/arXiv.1712.02034.

Groen, I.I., Greene, M.R., Baldassano, C., Fei-Fei, L., Beck, D.M., and Baker, C.I. (2018). Distinct contributions of functional and deep neural network features to representational similarity of scenes in human brain and behavior. Elife *7*, e32962. https://doi.org/10.7554/elife.32962.

Hamilton, W.L., Ying, Z., and Leskovec, J. (2017). Inductive representation learning on large graphs. In Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, pp. 1024–1034.

Hochreiter, S., and Schmidhuber, J. (1997). Long short-term memory. Neural Comput. *9*, 1735–1780. https://doi.org/10.1162/neco.1997.9.8.1735.

Hu, W., Liu, B., Gomes, J., Zitnik, M., Liang, P., Pande, V.S., and Leskovec, J. (2020). Strategies for pre-training graph neural networks. In 8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020.

Jain, A.K., Murty, M.N., and Flynn, P.J. (1999). Data clustering: a review. ACM Comput. Surv. *31*, 264–323. https://doi.org/10.1145/331499.331504.

Kho, R., Hansen, M., and Villar, H. (2006). Prevalence of Scaffolds in Human Cytochrome P450 Inhibitors Identified Using the Lopac1280 Library of Pharmacologically Active Compounds (Sigma-Aldrich). http://www.sigmaaldrich.com/Area_of_Interest/Life_Science/Life_Science_Quarterly/Spring_2006.html.

Kim, S., Thiessen, P.A., Bolton, E.E., Chen, J., Fu, G., Gindulyte, A., Han, L., He, J., He, S., Shoemaker, B.A., et al. (2016). Pubchem substance and compound databases. Nucleic Acids Res. *44*, D1202–D1213. https://doi.org/10.1093/nar/gkv951.

Kingma, D.P., and Ba, J. (2014). Adam: a method for stochastic optimization. Preprint at arXiv. https://doi.org/10.48550/arXiv.1412.6980.

Kipf, T.N., and Welling, M. (2016). Semi-supervised classification with graph convolutional networks. Preprint at arXiv. https://doi.org/10.48550/arXiv.1609.02907.

Landrum, G. (2006). Rdkit: open-source cheminformatics. https://github.com/rdkit/rdkit/releases/tag/Release_2016_09_4.

Lee, J.H., Basith, S., Cui, M., Kim, B., and Choi, S. (2017). In silico prediction of multiple-category classification model for cytochrome p450 inhibitors and non-inhibitors using machine-learning method. SAR QSAR Environ. Res. *28*, 863–874. https://doi.org/10.1080/1062936x.2017.1399925.

Li, X., and Fourches, D. (2020). Inductive transfer learning for molecular activity prediction: next-gen qsar models with molpmofit. J. Cheminf. *12*, 1–15. https://doi.org/10.26434/chemrxiv.9978743.v2.

Li, X., Grandvalet, Y., Davoine, F., Cheng, J., Cui, Y., Zhang, H., Belongie, S., Tsai, Y.H., and Yang, M.H. (2020). Transfer learning in computer vision tasks: remember where you come from. Image Vis Comput. *93*, 103853. https://doi.org/10.1016/j.imavis.2019.103853.

Li, X., Yan, X., Gu, Q., Zhou, H., Wu, D., and Xu, J. (2019). Deepchemstable: chemical stability prediction with an attention-based graph convolution network. J. Chem. Inf. Model. *59*, 1044–1049. https://doi.org/10.1021/acs.jcim.8b00672.

Lopez, S.A., Pyzer-Knapp, E.O., Simm, G.N., Lutzow, T., Li, K., Seress, L.R., Hachmann, J., and Aspuru-Guzik, A. (2016). The harvard organic photovoltaic dataset. Sci. Data *3*, 160086. https://doi.org/10.1038/sdata.2016.86.

Macarron, R., Banks, M.N., Bojanic, D., Burns, D.J., Cirovic, D.A., Garyantes, T., Green, D.V.S., Hertzberg, R.P., Janzen, W.P., Paslay, J.W., et al. (2011). Impact of high-throughput screening in biomedical research. Nat. Rev. Drug Discov. *10*, 188–195. https://doi.org/10.1038/nrd3368.

Mayr, A., Klambauer, G., Unterthiner, T., Steijaert, M., Wegner, J.K., Ceulemans, H., Clevert, D.A., and Hochreiter, S. (2018). Large-scale comparison of machine learning methods for drug target prediction on chembl. Chem. Sci. *9*, 5441–5451. https://doi.org/10.1039/C8SC00148K.

Mobley, D.L., and Guthrie, J.P. (2014). Freesolv: a database of experimental and calculated hydration free energies, with input files. J. Comput. Aided Mol. Des. *28*, 711–720. https://doi.org/10.1007/s10822-014-9747-x.

Papadatos, G., Gaulton, A., Hersey, A., and Overington, J.P. (2015). Activity, assay and target data curation and quality in the chembl database. J. Comput. Aided Mol. Des. *29*, 885–896. https://doi.org/10.1007/s10822-015-9860-5.

Paszke, A., Gross, S., Chintala, S., Chanan, G., Yang, E., DeVito, Z., Lin, Z., Desmaison, A., Antiga, L., and Lerer, A. (2017). Automatic differentiation in pytorch. In 31st Conference on Neural Information Processing Systems (NIPS 2017).

Ramakrishnan, R., Dral, P.O., Rupp, M., and von Lilienfeld, O.A. (2014). Quantum chemistry structures and properties of 134 kilo molecules. Sci. Data *1*, 140022. https://doi.org/10.1038/sdata.2014.22.

Ramsundar, B., Eastman, P., Walters, P., Pande, V., Leswing, K., and Wu, Z. (2019). Deep Learning for the Life Sciences (O'Reilly Media). https://www.amazon.com/Deep-Learning-Life-Sciences-Microscopy/dp/1492039837.

Ramsundar, B., Kearnes, S., Riley, P., Webster, D., Konerding, D., and Pande, V. (2015). Massively multitask networks for drug discovery. Preprint at arXiv. https://doi.org/10.48550/arXiv.1502.02072.

Rogers, D., and Hahn, M. (2010). Extended-connectivity fingerprints. J. Chem. Inf. Model. *50*, 742–754. https://doi.org/10.1021/ci100050t.

Rosenstein, M.T., Marx, Z., Kaelbling, L.P., and Dietterich, T.G. (2005). To transfer or not to transfer. In NIPS 2005 workshop on transfer learning, pp. 1–4.

Schneider, P., and Schneider, G. (2003). Collection of bioactive reference compounds for focused library design. QSAR Comb. Sci. *22*, 713–718. https://doi.org/10.1002/qsar.200330825.

Selzer, P., and Ertl, P. (2006). Applications of self-organizing neural networks in virtual screening and diversity selection. J. Chem. Inf. Model. *46*, 2319–2323. https://doi.org/10.1021/ci0600657.

Shen, J., and Nicolaou, C.A. (2020). Molecular property prediction: recent trends in the era of artificial intelligence. Drug Discov. Today Technol. *32–33*, 29–36. https://doi.org/10.1016/j.ddtec.2020.05.001.

Shrikumar, A., Greenside, P., Shcherbina, A., and Kundaje, A. (2016). Not just a black box: learning important features through propagating activation differences. Preprint at arXiv. https://doi.org/10.5555/3305890.3306006.

Simões, R.S., Maltarollo, V.G., Oliveira, P.R., and Honorio, K.M. (2018). Transfer and multi-task learning in qsar modeling: advances and challenges. Front. Pharmacol. *9*, 74. https://doi.org/10.3389/fphar.2018.00074.

Sterling, T., and Irwin, J.J. (2015). Zinc 15–ligand discovery for everyone. J. Chem. Inf. Model. *55*, 2324–2337. https://doi.org/10.1021/acs.jcim.5b00559.

Subramanian, G., Ramsundar, B., Pande, V., and Denny, R.A. (2016). Computational modeling of β-secretase 1 (bace-1) inhibitors using ligand based approaches. J. Chem. Inf. Model. *56*, 1936–1949. https://doi.org/10.1021/acs.jcim.6b00290.

Veith, H., Southall, N., Huang, R., James, T., Fayne, D., Artemenko, N., Shen, M., Inglese, J., Austin, C.P., Lloyd, D.G., and Auld, D.S. (2009). Comprehensive characterization of cytochrome p450 isozyme selectivity across chemical libraries. Nat. Biotechnol. *27*, 1050–1055. https://doi.org/10.1038/nbt.1581.

Veličković, P., Cucurull, G., Casanova, A., Romero, A., Lio, P., and Bengio, Y. (2017). Graph attention networks. Preprint at arXiv. https://doi.org/10.48550/arXiv.1710.10903.

Velickovic, P., Fedus, W., Hamilton, W.L., Liò, P., Bengio, Y., and Hjelm, R.D. (2019). Deep graph infomax. In 7th International Conference on Learning Representations, ICLR 2019. https://doi.org/10.48550/arXiv.1809.10341.

Wang, M., Zheng, D., Ye, Z., Gan, Q., Li, M., Song, X., Zhou, J., Ma, C., Yu, L., Gai, Y., et al. (2019a). Deep graph library: a graph-centric, highly-performant package for graph neural networks. Preprint at arXiv. https://doi.org/10.48550/arXiv.1909.01315.

Wang, Z., Dai, Z., Póczos, B., and Carbonell, J. (2019b). Characterizing and avoiding negative transfer. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 11293–11302. https://doi.org/10.1109/CVPR.2019.01155.

Weininger, D. (1988). Smiles, a chemical language and information system. 1. introduction to methodology and encoding rules. J. Chem. Inf. Model. *28*, 31–36. https://doi.org/10.1021/ci00057a005.

Williams, J.A., Hyland, R., Jones, B.C., Smith, D.A., Hurst, S., Goosen, T.C., Peterkin, V., Koup, J.R., and Ball, S.E. (2004). Drug-drug interactions for udp-glucuronosyltransferase substrates: a pharmacokinetic explanation for typically observed low exposure (auci/auc) ratios. Drug Metab. Dispos. *32*, 1201–1208. https://doi.org/10.1124/dmd.104.000794.

Xiong, Z., Wang, D., Liu, X., Zhong, F., Wan, X., Li, X., Li, Z., Luo, X., Chen, K., Jiang, H., and Zheng, M. (2019). Pushing the boundaries of molecular representation for drug discovery with the graph attention mechanism. J. Med. Chem. *63*, 8749–8760. https://doi.org/10.1021/acs.jmedchem.9b00959.

You, Y., Chen, T., Shen, Y., and Wang, Z. (2021). Graph contrastive learning automated. In Proceedings of the 38th International Conference on Machine Learning, ICML 2021 (PMLR), pp. 12121–12132. https://doi.org/10.48550/arXiv.2106.07594.

Zamir, A.R., Sax, A., Shen, W., Guibas, L.J., Malik, J., and Savarese, S. (2018). Taskonomy: disentangling task transfer learning. In Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 3712–3722. https://doi.org/10.1109/CVPR.2018.00391.

Zhang, W., Deng, L., and Wu, D. (2020). Overcoming negative transfer: a survey. Preprint at arXiv. https://doi.org/10.48550/arXiv.2009.00909.

Zhuang, F., Qi, Z., Duan, K., Xi, D., Zhu, Y., Zhu, H., Xiong, H., and He, Q. (2021). A comprehensive survey on transfer learning. Proc. IEEE *109*, 43–76. https://doi.org/10.1109/JPROC.2020.3004555.

STAR★METHODS

KEY RESOURCES TABLE

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Software and algorithms		
MoTSE	This study	https://github.com/lihan97/MoTSE
Python	Version 3.6.13	https://www.python.org/downloads/
PyTorch	Version 1.1.0	https://pytorch.org/
RDKit	Version 2018.09.3	https://www.rdkit.org/docs/Install.html
Deep Graph Library (DGL)	Version 0.4.2	https://www.dgl.ai/pages/start.html
Other		
QM9	(Ramakrishnan et al., 2014)	http://quantum-machine.org/datasets/
PCBA	(Ramsundar et al., 2015)	https://doi.org/10.48550/arXiv.1502.02072
Alchemy	(Chen et al., 2019)	https://www.dgl.ai/pages/start.html
FreeSolv	(Mobley and Guthrie, 2014)	https://alchemy.tencent.com/
BACE	(Subramanian et al., 2016)	https://doi.org/10.1021/acs.jcim.6b00290
HOPV	(Lopez et al., 2016)	https://doi.org/10.1038/sdata.2016.86

RESOURCE AVAILABILITY

Lead contact

Further information and requests for resources and reagents should be directed to and will be fulfilled by the lead contacts, Dan Zhao (zhaodan2018@tsinghua.edu.cn) and Jianyang Zeng (zengjy321@mail.tsinghua.edu.cn).

Materials availability

This study did not generate new unique reagents.

Data and code availability

- This paper analyzes existing, publicly available data. These accession numbers for the datasets are listed in the key resources table.
- The source code and datasets of MoTSE can be found at https://github.com/lihan97/MoTSE.
- Any additional information required to reanalyze the data reported in this paper is available from the lead contact upon request.

METHOD DETAILS

Notation and problem setting

Suppose that we are given a set of molecular property prediction tasks $\mathcal{T} = \{t_1, t_2, \ldots, t_N\}$, where N stands for the total number of tasks involved. Accordingly, we have a set of datasets $\mathcal{D} = \{D_1, D_2, \ldots, D_N\}$, where $D_i = \{(x, y)\}$ stands for the dataset related to task t_i, and (x, y) represents a pair of molecule and its label for task t_i. We represent each molecule as a graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$, where $\mathcal{V}$ stands for the set of nodes (i.e., heavy atoms) and $\mathcal{E}$ stands for the set of edges (i.e., covalent bonds). We use $u_k \in \mathbb{R}^{N_d}$ to represent the initial features (e.g., atom type) of the k-th node in $\mathcal{V}$, where N_d stands for the dimension of node features.

Our goal mainly lies in the following 2-folds: (1) efficiently estimate the similarity between each pair of tasks in $\mathcal{T}$; and (2) design an accurate transfer learning strategy based on the derived task similarity.

Key steps of MoTSE

Step 1: Pre-training the task-specific GNNs

The calculation of the similarity between tasks can be regarded as measuring the similarity of the intrinsic knowledge that needs to be learned from these tasks. Since deep learning models, especially the GNNs, have shown their superior capability of learning hidden knowledge and modeling various kinds of molecular properties (Gilmer et al., 2017; Li et al., 2019; Xiong et al., 2019), we adopt GNNs to capture such hidden knowledge contained in individual tasks. More specifically, for each task t, we pre-train a GNN model $m = p(e(\cdot))$ using the corresponding dataset D, where $e(\cdot)$ acts as an GNN encoder to extract the latent feature representations of the molecule graphs and $p(\cdot)$ serves as a classifier or regressor (implemented through a multi-layer perceptron) to make prediction for t.

Step 2: Projecting tasks into task space

After pre-training the task-specific GNNs for individual tasks, the problem of measuring the similarity between tasks is converted into finding a way to quantitatively represent the knowledge enclosed in the pre-trained GNNs. MoTSE employs two knowledge extraction methods, including an attribution method and a molecular representation similarity analysis (MRSA) method, to derive the hidden knowledge from GNN models as represented vectors in a latent space.

Before elaborating on the task projection methods, we first define a probe dataset $D_{probe} = \{x_1, x_2, \ldots, x_{N_p}\}$, which is a set of unlabeled molecules, where N_p denotes the number of molecules. This probe dataset is shared across all tasks involved and acts as a proxy in the knowledge extraction process of each task to ensure that all the tasks can be projected into a unified latent space.

Attribution method

The attribution method is a way of interpreting deep learning models by assigning importance scores for individual input features to explain the prediction. Here, we use an attribution method to assign importance scores to individual atoms in each molecule from the probe dataset. The specific attribution method we use is Gradient*Input (Shrikumar et al., 2016), which refers to a first-order Taylor approximation of how the output will change if a specific input feature is set to zero, thus indicating the importance of this input feature with respect to the output.

More formally, given the graph representation $\mathcal{G}$ for a molecule x from D_{probe}, the importance score a_k of the k-th atom u_k with respect to the task t can be computed as:

$$a_k = \frac{1}{N_d} \sum_{f=1}^{N_d} u_{k,f} \times \frac{\partial \hat{y}}{\partial u_{k,f}}, \qquad \text{(Equation 1)}$$

where $u_{k,f}$ stands for the f-th element of the feature vector u_k, N_d stands for the dimension of the input atom features, and $\hat{y} = m(\mathcal{G})$ stands for the prediction result of x for task t from the corresponding pre-trained GNN model m. Here, the importance score a_k of the k-th atom is derived by averaging the importance scores of all dimensions of the atom features. After assigning the importance scores to individual atoms of molecule x, we obtain an attribution vector, denoted by $A = [a_1, a_2, \ldots, a_{|\mathcal{V}|}]$, where $|\mathcal{V}|$ stands for the number of atoms in x. By applying the above attribution method to every molecule in D_{probe}, we can derive the attribution vectors of all molecules in the probe dataset, denoted by $\mathcal{A} = [A_1, A_2, \ldots, A_{N_p}]$.

Molecular representation similarity analysis

As the attribution method scores each atom separately without considering the global information of molecules, we define such knowledge extracted by the above attribution method as local knowledge. Here, we also present a molecular representation similarity analysis (MRSA) method (Groen et al., 2018; Dwivedi and Roig, 2019) to extract the global knowledge learned from the pre-trained GNNs. In particular, for each task, we compute the pairwise correlations between the hidden molecule representations (i.e., the outputs of the encoders of the pre-trained GNNs) to depict the relationships between molecules in the latent molecular representation space.

More formally, for a task t and the encoder e from the corresponding pre-trained model, we first perform forward propagation for all molecules in D_{probe} to generate their latent molecular representations

$Z = [z_1, z_2, \ldots, z_{N_p}]$, where z_m stands for the latent molecular representation of molecule $x_m \in D_{probe}$. Then for each pair of molecular representations z_m and z_n $(m \neq n)$, we compute their correlation score $r_{m,n}$, that is,

$$r_{m,n} = \rho(z_m, z_n), \quad \text{(Equation 2)}$$

where $\rho(\cdot)$ stands for the Pearson's correlation coefficient. After that, we obtain a molecular representation correlation vector $\mathcal{R} = [r_{1,2}, \ldots, r_{1,N_p}, r_{2,3}, \ldots, r_{N_p-1,N_p}]$ as another vector representation of task t.

For individual tasks in $\mathcal{T}$, MoTSE adopts the attribution method and the MRSA method mentioned above to extract both local and global task-related knowledge and projects them as vectors into two latent task spaces, denoted by $\mathcal{T}_{\mathcal{A}}$ and $\mathcal{T}_{\mathcal{R}}$, respectively.

Step 3: Estimating the task similarity

Once step 2 is completed, for each pair of tasks $t_i, t_j \in \mathcal{T}$ $(i \neq j)$, their similarity can be computed in the latent task spaces $\mathcal{T}_{\mathcal{A}}$ and $\mathcal{T}_{\mathcal{R}}$:

$$s^{\mathcal{A}}_{i,j} = \frac{1}{\mathcal{N}_p} \sum_{m=1}^{\mathcal{N}_p} cosine\_sim\left(\mathcal{A}^i_m, \mathcal{A}^j_m\right), \quad \text{(Equation 3)}$$

$$s^{\mathcal{R}}_{i,j} = cosine\_sim\left(\mathcal{R}^i, \mathcal{R}^j\right), \quad \text{(Equation 4)}$$

where $s^{\mathcal{A}}_{i,j}$ and $s^{\mathcal{R}}_{i,j}$ represent the task similarity derived in $\mathcal{T}_{\mathcal{A}}$ and $\mathcal{T}_{\mathcal{R}}$, respectively, and $cosine\_sim(\cdot)$ stands for the cosine similarity between two vectors.

The above two kinds of task similarities focus on different aspects to represent the hidden knowledge and are calculated under different assumptions. The attribution method mainly aims to extract local knowledge, and the assumption behind $s^{\mathcal{A}}_{i,j}$ is that similar tasks should have similar importance scores for the same atoms in a molecule. On the other hand, MRSA mainly aims to extract the global knowledge, and $s^{\mathcal{R}}_{i,j}$ measures the similarity on the basis that similar tasks should result in similar latent molecular representation spaces. To fully exploit the merits of both similarity estimation methods, we unify them into a more comprehensive formula:

$$s_{i,j} = (1 - \lambda) s^{\mathcal{A}}_{i,j} + \lambda s^{\mathcal{R}}_{i,j}, \quad \text{(Equation 5)}$$

where λ stands for the weighting factor.

The MoTSE-guided transfer learning strategy

After deriving the similarity between pairs of tasks in $\mathcal{T}$, for a target task $t_i \in \mathcal{T}$, we can select the task t_j $(i \neq j)$ with the highest similarity to t_i as the source task. As such, we can fine-tune the model pre-trained on dataset D_j to exploit the related knowledge from task t_j and thus enhance the prediction of target task t_i.

Implementation of MoTSE

Training details

The full network architecture for the GCN model is illustrated in Figure S2. We adopted a graph convolutional network (GCN) (Kipf and Welling, 2016) implemented by the deep graph library (DGL) (Wang et al., 2019a) as the encoder to model the molecular graphs and a two-layer perceptron as the predictor to make prediction for molecular properties. More specifically, the GCN encoder had three 256-dimensional GCN layers and the predictor was a two-layer (512-256-1) fully connected network. We employed weighted sum pooling and max pooling as readout functions to produce the global feature representations of molecules and used a concatenation operation to combine these two derived feature representations as the final molecular feature representation. We employed ReLU as the activation function and set the dropout rate to zero. The pre-training and fine-tuning shared the same set of hyper-parameters. All the models were trained with the Pytorch framework (Paszke et al., 2017). The MSELoss and CrossEntropyLoss functions were employed to measure the mean-squared error and the cross entropy for the regression tasks and classification tasks, respectively. We used the Adam optimizer (Kingma and Ba, 2014) for gradient descent optimization with the following hyper-parameters: learning rate 1×10^{-4} and weight decay 1×10^{-5}. All the models were trained for 200 epochs with early stopping, which aimed to terminate training when the validation accuracy had not been improved in the last 20 epochs. As MoTSE and baseline learning strategies

are orthogonal to different model architectures, we did not tune the model configurations. The model and the training configurations of MoTSE were shared with baseline learning strategies for a fair comparison.

The probe dataset

In our tests, we constructed a probe dataset by randomly sampling 500 small molecules from the ZINC dataset (Sterling and Irwin, 2015). Although a larger probe dataset with carefully selected molecules may serve as a better proxy in the knowledge extraction process, intuitively, the empirical results demonstrated that 500 randomly selected molecules were sufficient to provide reliable estimations. More details about the effects of the randomness and the size of the probe dataset on the performance of MoTSE are provided in supplementary section 1.1.

Task similarity estimation

We empirically set the weighting factor λ to 0.7 in our computational experiments (see supplementary section 1.2 for more details).

Datasets and data processing

Datasets used to evaluate the prediction performance

We mainly used four representative datasets, including QM9 (Ramakrishnan et al., 2014), PCBA (Ramsundar et al., 2015), FreeSolv (Mobley and Guthrie, 2014), BACE (Subramanian et al., 2016), HOPV (Lopez et al., 2016) and Alchemy (Chen et al., 2019), to evaluate the effectiveness of our proposed method in molecular property prediction. QM9 is a dataset that provides twelve quantum chemical properties, such as geometric, energetic, electronic and thermodynamic properties of roughly 130K small molecules, associated with twelve regression tasks (Ramakrishnan et al., 2014). PCBA is a dataset consisting of biological activities of small molecules generated by high-throughput screening, associated with 128 classification tasks (Ramsundar et al., 2015). The FreeSolv dataset measures the hydration free energy of 642 small molecules in water from both experiments and alchemical free energy calculation (Mobley and Guthrie, 2014). The BACE dataset measures whether each of 1,513 molecules can act as an inhibitor of human β-secretase 1 (BACE-1) (Subramanian et al., 2016). HOPV is a dataset that provides eight quantum chemical properties containing 350 organic donor compounds (Lopez et al., 2016). The Alchemy dataset (Chen et al., 2019) shares the same tasks as the QM9 dataset but has different data distributions, that is, the QM9 dataset contains molecules comprising up to nine non-hydrogen atoms while the molecules in the Alchemy dataset consist of nine to fourteen non-hydrogen atoms.

We summarized the details of the tasks for the preprocessed QM9 and PCBA datasets, the FreeSolv dataset, the BACE dataset, the HOPV dataset and the Alchemy dataset in Table S1.

Dataset measuring physical chemistry properties

We constructed a dataset containing 10K molecules labeled with four physical properties, including counts of N and O atoms (NOCount), counts of NH and OH atoms (NHOHCount), number of H acceptors (NHA) and number of H donors (NHD). More specifically, We randomly sampled 10K molecules from the ZINC dataset (Sterling and Irwin, 2015) and derived the four properties from RDKit (Landrum, 2006).

Dataset measuring the bio-activities against cytochrome P450 isozymes

We also obtained a dataset that estimates the bio-activities of 17K molecules against five cytochrome P450 isozymes, including 1A2, 2C9, 2C19, 2D6 and 3A4, from a preprocessed ChEMBL dataset (Mayr et al., 2018).

Learning strategies

The task similarity derived from MoTSE was employed to guide the source task selection in transfer learning. More specifically, for each target task in $QM9_{1k}$/$PCBA_{1k}$, we selected n tasks with the top similarity scores as the source tasks from $QM9_{10k}$/$PCBA_{10k}$ according to the task similarity estimated by MoTSE, and took the best fine-tuning results as the final results. We set n to three and five for QM9 and PCBA datasets, respectively. The effect on the choice of n is provided in supplementary section 1.2.

To benchmark our proposed transfer learning strategy, we employed various previously defined transfer learning strategies, which mainly differed in the ways of defining source tasks and leveraging the knowledge from source tasks. More specifically, we employed multitask learning (denoted as MT), which learned

the target task and all the available source tasks simultaneously, five self-supervised learning methods, including Masking (Hu et al., 2020), EdgePred (Hamilton et al., 2017), ContextPred (Hu et al., 2020), DeepGraphInfomax (Velickovic et al., 2019) (denoted as DGI) and JOAO (You et al., 2021), which first leveraged different proxy tasks to learn general knowledge from a large-scale unlabeled dataset and then fine-tuned the pre-trained model on the target dataset (here we used the ZINC dataset (Sterling and Irwin, 2015) with two million molecules to pre-train the self-supervised learning methods), and another four self-supervised learning methods, including EdgePred_{sup} (Hu et al., 2020), Masking_{sup} (Hu et al., 2020), ContextPred_{sup} (Hu et al., 2020) and DGI_{sup} (Hu et al., 2020), which first pre-trained the models using self-supervised strategies, then further pre-trained them on a preprocessed ChEMBL dataset (Mayr et al., 2018) by learning to predict bio-activities in a supervised fashion and finally fine-tuned the pre-trained models on target datasets. Moreover, we introduced the training from scratch scheme (denoted as Scratch) as a baseline method, which directly trained the model on the dataset of the target task and did not exploit any extra knowledge in the learning process.

Note that, our method is orthogonal to different GNN architectures. Here, for a fair comparison, we implemented all the learning strategies on the basis of GCNs if not specially specified, and we also used the same set of hyper-parameters for each method.

Model configurations

To evaluate whether the similarity estimated using GCNs can be generalized to guide the source task selection of other model architectures, we also considered other models, including graph attention networks (GATs) (Veličković et al., 2017), fully-connected networks (FCNs) and recurrent neural networks (RNNs) in our tests. The details of the models are provided below.

- GAT: GAT is a kind of graph neural network that employs the attention mechanism when performing message passing over nodes. We constructed a GAT model with three 256-dimensional GAT layers followed by a two-layer (512-256-1) fully connected network with the ReLU activation function (Agarap, 2018) for molecular representation extraction and property prediction.
- FCN: We built a five-layer (2048-1024-512-256-128-1) fully connected network with the ReLU activation function (Agarap, 2018) that took the ECFP (extended connectivity fingerprints) (Rogers and Hahn, 2010) representations of molecules as input to make molecular property prediction.
- RNN: RNN is a deep learning model particularly designed for processing sequential data, which has been proven to be effective in making molecular property prediction with SMILES (simplified molecular input line entry specification) representations (Weininger, 1988; Goh et al., 2018; Arús-Pous et al., 2019). Here, we employed a three-layer 128-dimensional LSTM (Hochreiter and Schmidhuber, 1997) (a classical variant of RNN) to encode SMILES representations of molecules into 64-dimensional latent vectors and a two-layer (64-32-1) fully connected neural network with the ReLU activation function (Agarap, 2018) to make predictions.

ECFP is a representation of 2D binary fingerprints (i.e., a series of bits) which can dynamically index the presence or absence of particular substructures of molecules. SMILES is a string-based molecular representation for describing molecular structures using short ASCII strings. Figure S1 gives an example for the ECFP and SMILES representations. In our computational experiments, we used DeepChem (Ramsundar et al., 2019) to calculate a 2048-bit ECFP for each molecule. For the SMILES string, we used one-hot vectors to encode the unique characters.