Scalable Multi-Task Transfer Learning for Molecular Property Prediction

Chanhui Lee [1 2] **Dae-Woong Jeong** [1] **Sung Moon Ko** [1] **Sumin Lee** [1] **Hyunseung Kim** [1] **Soorin Yim** [1] **Sehui Han** [1] **Sungwoong Kim** [2] **Sungbin Lim** [1 3]

Abstract

Molecules have a number of distinct properties whose importance and application vary. Often, in reality, labels for some properties are hard to achieve despite their practical importance. A common solution to such data scarcity is to use models of good generalization with transfer learning. This involves domain experts for designing source and target tasks whose features are shared. However, this approach has limitations: i). Difficulty in accurate design of source-target task pairs due to the large number of tasks, and ii). corresponding computational burden verifying many trials and errors of transfer learning design, thereby iii). constraining the potential of foundation modeling of multi-task molecular property prediction. We address the limitations of the manual design of transfer learning via data-driven bi-level optimization. The proposed method enables scalable multi-task transfer learning for molecular property prediction by automatically obtaining the optimal transfer ratios. Empirically, the proposed method improved the prediction performance of 40 molecular properties and accelerated training convergence.

1. Introduction

Given that a molecule has a number of molecular properties, basically, molecular property prediction is to predict a target property among various properties (Wieder et al., 2020). There are many important applications of molecular property prediction, including virtual screening and discovery of novel materials and drugs (Christopher et al., 2001; Atanasov et al., 2021; Gentile et al., 2022; Sadybekov & Katritch, 2023). Molecular property prediction provides vital information in making informed decisions throughout the discovery and development process so that the development cycle of the product can be accelerated.

However, in reality, molecular property prediction often suffers from the data scarcity problem (Hu et al., 2019; Li et al., 2022; 2021) due to various factors, including the high cost of experimental data generation (Axelrod & Gómez-Bombarelli, 2022), the complexity of chemical compounds (Kumar et al., 2020), and the proprietary nature of pharmaceutical data (Heyndrickx et al., 2023). As a result, researchers have tried to overcome this limitation, such as using advanced computational techniques like transfer learning (Ko et al., 2024b), data augmentation (You et al., 2020), and other approaches (Lu et al., 2019; Yao et al., 2023; Qian et al., 2023) that can learn effectively from smaller datasets.

Transfer learning (Pan & Yang, 2010; Zhuang et al., 2019) allows for the effective generalization of knowledge learned from source task data distribution to the target task data. Effective transfer can be fulfilled through learning mutually informative feature representations between aligned tasks. GATE (Ko et al., 2024b) introduced a geometric alignment of various tasks to enhance task alignment for molecular property prediction. With the proposed geometrical alignment, the prediction model can learn geometrically aligned molecular representations that are applicable from source task to target task, enabling effective transfer learning and surpassing the performance of baseline models.

Recently, Ko et al. (2024a) extended GATE to a multi-task setting by sharing a single latent space among multiple tasks, applying geometrical alignment regularization within this shared latent space. In this extended formulation of GATE, different transfer ratios can be applied for each (source, target) task pair, which represent a belief in how much a source task can be helpful to the target task. In Figure 1, we investigated the effect of different transfer ratios for three molecular properties: density (ds), heat of vaporization (hv), and boiling point (bp). Assuming the same transfer ratio between paired tasks, we conducted a grid search over [0.2, 1.0] for 3 hyperparameters $\lambda_{hv,bp}$, $\lambda_{ds,bp}$, $\lambda_{hv,ds}$ and found that overall prediction performance largely varies by the setting of transfer ratios.

Though the transfer ratio could have a large effect on the prediction performance, GATE lacks a method for explor-

[1] LG AI Research, Seoul, Republic of Korea [2] Department of Artificial Intelligence, Korea University, Seoul, Republic of Korea [3] Department of Statistics, Korea University, Seoul, Republic of Korea. Correspondence to: Sungbin Lim <sungbin@korea.ac.kr>.

AI4Science Workshop of the 40^{th} International Conference on Machine Learning, Copyright 2024 by the author(s)

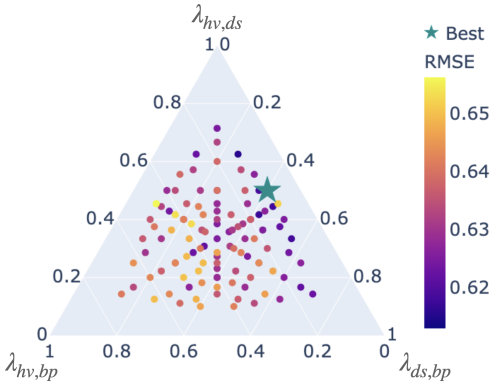

Figure 1. Grid search of transfer ratios λ between density (ds), heat of vaporization (hv), and boiling point (bp). Each axis $\lambda_{hv,bp}, \lambda_{ds,bp}, \lambda_{hv,ds}$ corresponds to the transfer ratio between (hv, bp), (ds, bp), (hv, ds), assuming $\lambda_{i \to j} = \lambda_{j \to i}$. The color of a point corresponds to the Root Mean Square Error (RMSE) of model prediction at the end of training with $\lambda_{hv,bp}, \lambda_{ds,bp}, \lambda_{hv,ds}$, and the best hyperparameter set is marked as a star.

ing transfer ratios but rather uses them as hyperparameters, which presents several limitations. i. Inaccuracy in predicted transfer ratios by domain experts: Given the black-box nature of deep learning models, there is no guarantee that the source task and the corresponding source data chosen by a domain expert will actually enhance the performance of the target task. ii. Limited scalability: As the number of tasks increases, manually setting proper ratios for all possible transfer interactions between two tasks by a domain expert becomes infeasible and less optimized. iii. Constraining the foundation modeling in molecular property prediction: As different tasks in molecular property prediction fundamentally involve comprehending molecular structures, the performance of tasks with limited data can be improved by merging existing datasets for extensive multi-task training. This approach maximizes the benefits of foundational modeling in predicting molecular properties.

To address these limitations, we propose a novel bi-level optimization method to automatically obtain the optimal transfer ratios for given multi-task data. This bi-level optimization replaces previous manual hyperparameter searches by domain experts through gradient-based learning on the validation performance. The training algorithm remains the same during the training phase; the difference occurs in the validation phase. In the validation phase, gradients flow from the computed loss to the computation node representing the transfer ratio and are updated gradient-based on their contribution to the loss value during the validation phase. Since the gradient computation is restricted to the transfer ratio, additional time and space costs for the proposed bi-level optimization are negligible. The proposed gradient-based bi-level optimization efficiently obtains the optimal transfer ratio, especially on a large task space, without cumbersome tuning by domain experts.

Contribution

- We propose a data-driven method to search optimal transfer ratios for multi-task transfer learning of molecular property prediction.
- The proposed method has improved the performances on 40 tasks of molecular property regression.
- The proposed method accelerates the convergence of multi-task transfer learning in molecular property regressions.

2. Preliminary: Multi-task property regression

This section introduces preliminary works for multi-task property regression. For multi-task transfer learning in property regression, we leverage GATE algorithms extended for multi-task transfer learning (Ko et al., 2024a).

2.1. Multi-task learning extension of GATE

GATE addresses transfer learning among a number of tasks, introducing additional side tasks to learn mutually useful features for different tasks in shared manifold $\mathcal{M}$. $\mathcal{M}$ is a manifold where each task-specific model can learn the general geometrical knowledge of molecular structure. This strategy guides the model in learning generally useful features for molecular property regression, allowing the task of scarce data to take advantage of knowledge learned from another data-enriched task.

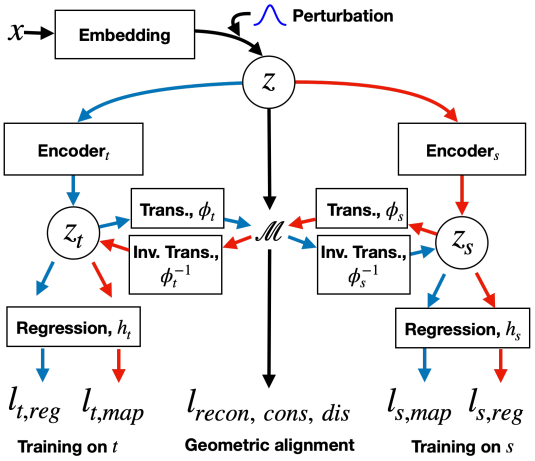

Figure 2. Training overview of GATE. t, s represents the target and source tasks for transfer learning. The colors of the arrows differentiate prediction paths: red corresponds to the path from Encoder$_s$, and blue corresponds to the path from Encoder$_t$.

2.2. Target task regression

When a molecule x is fed into an embedding model as SMILES (Weininger, 1988), we get a corresponding embedding vector z. Given a target task t, a task-specific encoder (encoder$_t$) embeds z into a latent vector z_t on the manifold for the target task t. Then, the regression head h_t predicts $\hat{y}_t$ to calculate the Mean Squared Error (MSE) loss with respect to target label y_t.

$$z_t = \text{encoder}_t(z) \tag{1}$$

$$\hat{y}_t = h_t(z_t) \tag{2}$$

$$l_{\text{reg}} = \frac{1}{N}\sum_{i}^{N} \text{MSE}(y_t, \hat{y}_t) \tag{3}$$

where N is the number of data points.

2.3. Transfer learning from source task

Let s be another task we can leverage for target task learning. Thanks to the shared manifold $\mathcal{M}$, z_t can be represented from a source task representation z_s, via transformation $\phi_{s\to\mathcal{M}}$ and inverse transformation $\phi^{-1}_{\mathcal{M}\to t}$.

$$z_s = \text{encoder}_s(z) \tag{4}$$

$$z_{\mathcal{M}} = \phi_{s\to\mathcal{M}}(z_s) \tag{5}$$

$$z_t = \phi^{-1}_{\mathcal{M}\to t}(z_{\mathcal{M}}) \tag{6}$$

$$\hat{y}_t = h_t(z_t) \tag{7}$$

where $\phi_{s\to\mathcal{M}}(z_s)$ means a transformation of the vector from the manifold of source task s to shared manifold $\mathcal{M}$, and $\phi^{-1}_{\mathcal{M}\to t}(z_s)$ means an inverse transformation from the shared manifold $\mathcal{M}$ to the manifold of target task t.

With a hyperparameter called a mapping ratio $\lambda_{s\to t}$, GATE conducts transfer learning from source task s to target task t.

$$l_{map} = \sum_{s} \frac{1}{N}\sum_{i}^{N} \lambda_{s\to t}\text{MSE}(y_t, \hat{y}_t) \tag{8}$$

As the correlation varies across different source-target task pairs, the effectiveness of multi-task transfer learning is dependent on the proper search of the $\lambda_{s\to t}$. For instance, the Highest Occupied Molecular Orbital (HOMO) and Lowest Unoccupied Molecular Orbital (LUMO) tasks would be deeply correlated, sharing many of the necessary features representing molecular orbitals. Therefore, a high $\lambda_{s\to t}$ value can accelerate the mutual learning of the HOMO and LUMO. However, in a multi-task learning setting, the correlation of the target-source task is conditioned on other tasks, which makes it more difficult to find optimal λ with many tasks to learn. In this situation, completing the entire correlation of target-source task pairs is prohibitive even for experienced domain experts.

2.4. Geometric regularizations

To align manifolds of different tasks to be the shared manifold $\mathcal{M}$, GATE aims to align the geometric representation of a molecule in different properties through side tasks: reconstruction, consistency, and distance.

Reconstruction To convince those models can learn general geometries useful across tasks, $\mathcal{M}$ should have enough expressiveness to reconstruct z. The following reconstruction loss

$$l_{\text{ae}} = \sum_{i} \text{MSE}(z_i, \phi_{\mathcal{M}\to i}(\phi^{-1}_{i\to\mathcal{M}}(z_i)), \tag{9}$$

regularizes GATE to maintain the reconstruction capability.

Consistency Too much divergence between the manifold of t, s could harm the transfer effect. To regularize the significant divergence between z_t, z_s on the shared manifold $\mathcal{M}$, GATE introduces a loss for consistency,

$$l_{cons} = \sum_{s} \text{MSE}(z_s, z_t). \tag{10}$$

Distance To learn robust transformation $\phi_{i\to\mathcal{M}}, \phi^{-1}_{\mathcal{M}\to i}$ under perturbation, GATE applies a perturbation on z to have z'. Then regularizes the resulting displacement for different tasks i to be minimized as

$$l_{dis} = \frac{1}{M}\sum_{s} C_s \sum_{p=1}^{M} \text{MSE}(s_s^p, s_t^p), \tag{11}$$

where M is the number of perturbed points, superscript p means p-th perturbed point, C$_s$ is the hyperparameter, and $s_s^p = \|\phi_{i\to\mathcal{M}}(\text{encoder}_i(z)) - \phi_{i\to\mathcal{M}}(\text{encoder}_i(z^p))\|$.

Aggregating the losses for geometrical alignments, the total loss is calculated as

$$l_{tot} = l_{reg} + l_{ae} + l_{cons} + l_{dis} + l_{map}, \tag{12}$$

to update model parameters for ϕ, ϕ^{-1}, h, and the embedding model. For brevity, we simply represent the corresponding model parameters as θ.

3. Bi-level optimization of GATE

We interpret the problem of finding optimal λ as a bi-level optimization problem:

$$\begin{aligned} \min_{\lambda} \quad & l_{\text{val}}(\theta, \lambda) \\ \text{s.t.} \quad & \theta^*(\lambda) = \arg\min_{\theta} l_{\text{train}}(\theta, \lambda), \end{aligned} \tag{13}$$

Algorithm 1 depicts the detailed algorithms for the bi-level optimization of GATE. First, θ is updated using the training dataset D_{tr} and λ in the inner loop, and in the outer loop, based on the updated θ, λ is updated with respect to the performance in the validation dataset D_{val}.

Algorithm 1 Bi-level Optimization for GATE

```
**Input:** Training data $D_{tr}$, validation data $D_{val}$
Initialize model $\theta$, transfer ratio $\lambda$, transfer momentum $m, v$
**repeat**
    $\theta \leftarrow \arg\min_{\theta} L(D_{tr}, \theta, \lambda)$ {Inner loop}
    $\lambda \leftarrow \arg\min_{\lambda} L(D_{val}, \theta, \lambda, m, v)$ {Outer loop}
**until** converged =0
```

In the inner loop, our objective is to update θ given λ and D_{tr}. Given a molecule, x, the embedding model projects x into the embedding vector z. Subsequently, encoder$_i$ projects z to z_i for all tasks $i \in T \cup S$, where T, S represent sets of target tasks and source tasks, respectively. Then, we can get two predictions for the target property t; one is directly from z_t and the other is from $\phi^{-1}_{\mathcal{M}\to t}(\phi_{s\to\mathcal{M}}(z_s))$. The predicted values are used to calculate the regression loss l_{reg} and mapping loss l_{map}, respectively. After summation of losses l_{reg} and l_{map} with $l_{\text{ae}}, l_{\text{cons}}, l_{\text{dis}}$, finally θ is updated from the corresponding gradient.

Algorithm 2 Inner loop

```
**Input:** Target tasks $T = \{t_1, \cdots, t_{N_T}\}$, Source tasks $S = \{s_1, \cdots, s_{N_S}\}$, model parameters $\theta$, transfer ratios $\lambda$, training data $(x_i, y_i)$ $\forall i \in T \cup S$
**Output:** Optimized model parameters $\theta$
**while** Training epoch **do**
    $l_{tot} = 0$
    $z = \text{embedding}(x),\ \forall i \in T \cup S$
    $z_i = \text{encoder}_i(z),\ \forall i \in T \cup S$
    **for** $(t, s)$ **in** $\{(t, s) | t \in T, s \in S\}$ **do**
        $l_{reg} = \text{MSE}(y_t, h_t(z_t))$
        $l_{map} = \lambda_{s\to t}\text{MSE}(y_t, h_t(\phi^{-1}_{\mathcal{M}\to t}(\phi_{s\to\mathcal{M}}(z_s))))$
        $l_{tot} \mathrel{+}= l_{reg} + l_{map} + l_{ae} + l_{cons} + l_{dis}$
    **end for**
    $\theta \leftarrow \nabla_\theta l_{tot}$
**end while**=0
```

In the outer loop, we aim to search for $\lambda = \min_{\lambda} L_{\text{val}}(\theta^*, \lambda)$. The mapping loss between the target and source task is calculated to get gradient g with respect to the updated θ through the inner loop. Then g is used to update the moving average m and the squared moving average v, with the corresponding hyperparameters $\beta_0, \beta_1 \in [0, 1)$. Finally, the bias-corrected estimate of the first moment and the second moment, $\hat{m}$ and $\hat{v}$, are calculated to update the transfer ratio λ. Markedly, the calculated mapping loss in the outer loop only updates λ without affecting θ. This procedure is a data-driven hyperparameter search that substitutes for manual hyperparameter search, which is a bottleneck for the foundation modeling of multi-task transfer learning for molecular property regression.

Algorithm 3 Outer loop

```
**Input:** Target tasks $T = \{t_1, \cdots, t_{N_T}\}$, Source tasks $S = \{s_1, \cdots, s_{N_S}\}$, model parameters $\theta$, transfer ratio $\lambda$, transfer momentum $m, v$, validation data $(x_i, y_i)$ $\forall i \in T \cup S$
**Output:** Optimized $\lambda$
**while** Validation epoch **do**
    step $\leftarrow$ step $+ 1$
    $l_{map} = 0$
    $z = \text{embedding}(x),\ \forall i \in T \cup S$
    $z_i = \text{encoder}_i(z),\ \forall i \in T \cup S$
    **for** $(t, s)$ **in** $\{(t, s) | t \in T, s \in S\}$ **do**
        $l_{map} \mathrel{+}= \lambda_s \text{MSE}(y_t, h_t(\phi^{-1}_{\mathcal{M}\to t}(\phi_{s\to\mathcal{M}}(z_s))))$
    **end for**
    $g \leftarrow \nabla_\lambda l_{map}$
    $m \leftarrow \beta_0 m + (1 - \beta_0) g,\ v \leftarrow \beta_1 v + (1 - \beta_1) g^2$
    $\hat{m} \leftarrow \frac{m}{1-\beta_0^{\textbf{step}}},\ \hat{v} \leftarrow \frac{v}{1-\beta_1^{\textbf{step}}}$
    $\lambda \leftarrow \lambda - \eta \frac{\hat{m}}{\sqrt{\hat{v}}+\epsilon}$
**end while**=0
```

4. Experiments

4.1. Dataset

To test in a scaled-up multi-task setting, we collected data from 40 tasks from PubChem (Kim et al., 2022), Ochem (Sushko et al., 2011), CCDDS, Yaws Handbook, and Jean-Claude Bradley. We specified each task and number of data points in Appendix A. For the robust test, we used scaffold split of the train and test dataset based on the molecular structure (Bemis & Murcko, 1996). To avoid the overfitting of transfer ratio adaptation to the validation dataset, we interchanged 20% of the train and validation datasets for every epoch.

4.2. Results

Table 1 shows the performance of GATE with and without bi-level optimization on 40 different molecular property regression tasks. To evaluate the effectiveness of the proposed method, other than λ, we used the same model architecture and hyperparameters of GATE. Performance is measured in terms of Root Mean Square Error (RMSE), a standard metric used to measure prediction accuracy. A lower RMSE indicates better performance. The results show that GATE with bi-level optimization generally achieves lower RMSE scores across most tasks than the original GATE method. This is achieved by learning the transfer ratio λ to minimize prediction error described in Equation (8), which

Table 1. 40 task RMSE of GATE and GATE integrated with bi-level optimization (GATE*). The best case for each task is highlighted in bold.

	Tasks									
Method	CPS	LM	UFT	VS	MAS	CTP	GEF	PKA	MVS	DM
GATE	0.154	0.367	0.497	**0.650**	**0.343**	0.402	0.326	0.765	0.250	0.690
GATE*	**0.139**	**0.366**	**0.475**	0.739	0.376	**0.373**	**0.236**	**0.731**	**0.233**	**0.687**
	SEF	CPL	HVC	FP	PAR	CT	VP	LP	CPG	IP
GATE	0.452	0.252	0.562	0.645	**0.241**	0.414	0.730	**0.134**	**0.119**	**0.501**
GATE*	**0.358**	**0.133**	**0.555**	**0.636**	0.281	**0.408**	**0.685**	0.135	0.134	0.533
	SPA	LF	ST	HF	RI	AS	DS	NEC	DK	MVL
GATE	0.259	**0.510**	**0.409**	0.561	0.443	**0.754**	0.477	0.323	0.591	0.484
GATE*	**0.234**	0.512	0.420	**0.516**	**0.418**	0.807	**0.442**	**0.257**	**0.559**	**0.444**
	HV	SAE	BP	CTV	MP	HC	POL	HM	AW	ROG
GATE	0.478	0.263	0.520	0.284	0.583	0.265	0.307	0.441	0.711	0.552
GATE*	**0.431**	**0.258**	**0.508**	**0.237**	**0.579**	**0.246**	**0.285**	**0.432**	**0.596**	**0.548**

enables more effective transfer learning than using constant λ, though given the same data for target task and source tasks. Specifically, as Section 4.2, the performances were enhanced in 31 out of 40 tasks, reducing the average RMSE by 4.4%. This suggests that incorporating bi-level optimization in GATE improves prediction accuracy across a wide range of tasks.

Table 2. 40 task RMSE Improvements by applying bi-level optimization over vanilla GATE

Method	No. Improved Tasks	Avg RMSE
GATE	-	100%
GATE*	31	95.6%

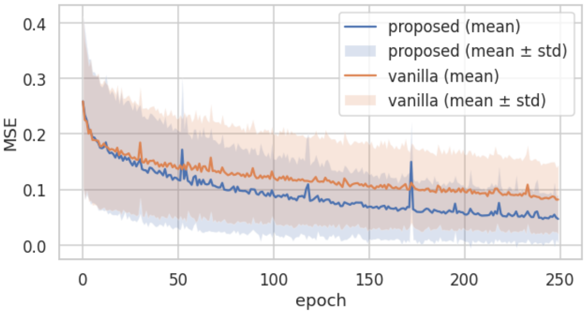

Figure 3. Validation loss curve in learning 40 tasks molecular property regression, with and without the proposed methods.

In addition, we found that applying bi-level optimization not only enhances performance but also accelerates loss convergence. Figure 3 shows that with the same training epoch, regression loss converges much faster with the proposed method. The fast convergence is due to the learning of $\lambda_{s\rightarrow t}$, which strengthens the transfer between highly correlated molecular properties and, at the same time, controls too much transfer between less correlated molecular properties. In the end, the variance across the different tasks is reduced, as shown by the narrower shade of the proposed method than vanilla GATE.

5. Discussion

Algorithm 3 imply that we can accelerate the outer loop with reduced GPU memory usage by only backpropagating the gradient of tasks whose λ is above a threshold. Training 40 molecular property prediction tasks, we found that this direction is promising, as the 95% of quadratic variations of λ are under 0.1. This means that updates from many source and target task pairs do not result in a big update of the model parameters. We hope this direction guides future works to improve the time and space complexity of the proposed method.

6. Conclusion

This study presents a bi-level optimization approach for enhancing transfer learning in multi-task property regression on a large scale. The performance in multi-task transfer learning is significantly influenced by how the correlation between the source and target tasks is modeled. Typically, designing this correlation has relied on domain experts. However, with increasing tasks, relying solely on domain experts for correlation design needs impractical time and inaccurate design due to the exponentially increasing number of task pair combinations, which can lead to sub-optimal outcomes. To address this issue, we employ a data-driven bi-level optimization strategy to identify the optimal correlation design. In our evaluation across 40 tasks, applying our method decreased RMSE for 31 tasks, with an average reduction of 4.4%.

References

Atanasov, A. G., Zotchev, S. B., Dirsch, V. M., Bay, I. E. M. J. M. D. W. R. E. O. B. R. B. W. B., Orhan, I. E., Banach, M., Rollinger, J. M., Barreca, D., Weckwerth, W., Bauer, R., Bayer, E. A., Majeed, M., Bishayee, A., Bochkov, V. N., Bonn, G. K., Braidy, N., Bucar, F., Cifuentes, A., D'Onofrio, G., Bodkin, M. A., Diederich, M. F., Dinkova-Kostova, A. T., Efferth, T., Bairi, K. E., Arkells, N., Fan, T.-P., Fiebich, B. L., Freissmuth, M., Georgiev, M. I., Gibbons, S., Godfrey, K. M., Gruber, C. W., Heer, J. P., Huber, L. A., Ibáñez, E., Kijjoa, A., Kiss, A. K., Lu, A., Macias, F. A., Miller, M. J. S., Mocan, A., Müller, R., Nicoletti, F., Perry, G., Pittalà, V., Rastrelli, L., Ristow, M., Russo, G. L., Silva, A. S., Schuster, D., Sheridan, H., Skalicka-Woźniak, K., Skaltsounis, L. A., Sobarzo-Sánchez, E., Bredt, D. S., Stuppner, H., Sureda, A., Tzvetkov, N. T., Vacca, R. A., Aggarwal, B. B., Battino, M. A., Giampieri, F., Wink, M., Wolfender, J., Xiao, J., Yeung, A. W. K., Lizard, G., Popp, M. A., Heinrich, M., Berindan-Neagoe, I., Stadler, M., Daglia, M., Verpoorte, R., and Supuran, C. T. Natural products in drug discovery: advances and opportunities. *Nature Reviews. Drug Discovery*, 20:200 – 216, 2021. URL https://api.semanticscholar.org/CorpusID:231726688.

Axelrod, S. and Gómez-Bombarelli, R. Geom, energy-annotated molecular conformations for property prediction and molecular generation. *Scientific Data*, 9, 2022. URL https://api.semanticscholar.org/CorpusID:256834165.

Bemis, G. W. and Murcko, M. A. The properties of known drugs. 1. molecular frameworks. *Journal of medicinal chemistry*, 39(15):2887–2893, 1996.

Christopher, Lipinski, A., Beryl, Dominy, W., Paul, and Feeney, J. Experimental and computational approaches to estimate solubility and permeability in drug discovery and development settings. *Advanced drug delivery reviews*, 46 1-3:3–26, 2001. URL https://api.semanticscholar.org/CorpusID:24301532.

Gentile, F., Yaacoub, J. C., Gleave, J., Fernández, M., Ton, A.-T., Ban, F., Stern, A., and Cherkasov, A. Artificial intelligence–enabled virtual screening of ultra-large chemical libraries with deep docking. *Nature Protocols*, 17:672 – 697, 2022. URL https://api.semanticscholar.org/CorpusID:246556999.

Heyndrickx, W., Mervin, L. H., Morawietz, T., Sturm, N., Friedrich, L., Zalewski, A., Pentina, A., Humbeck, L., Oldenhof, M., Niwayama, R., Schmidtke, P., Fechner, N., Simm, J., Arany, A., Drizard, N., Jabal, R., Afanasyeva, A., Loeb, R., Verma, S., Harnqvist, S., Holmes, M., Pejó, B., Teleńczuk, M., Holway, N., Dieckmann, A., Rieke, N., Zumsande, F., Clevert, D.-A., Krug, M., Luscombe, C. N., Green, D. V. S., Ertl, P., Antal, P., Marcus, D., Huu, N. D., Fuji, H., Pickett, S., Ács, G., Boniface, E., Beck, B., Sun, Y., Gohier, A., Rippmann, F., Engkvist, O., Göller, A. H., Moreau, Y., Galtier, M., Schuffenhauer, A., and Ceulemans, H. Melloddy: Cross-pharma federated learning at unprecedented scale unlocks benefits in qsar without compromising proprietary information. *Journal of Chemical Information and Modeling*, 64:2331 – 2344, 2023. URL https://api.semanticscholar.org/CorpusID:261339793.

Hu, W., Liu, B., Gomes, J., Zitnik, M., Liang, P., Pande, V. S., and Leskovec, J. Pre-training graph neural networks. *ArXiv*, abs/1905.12265, 2019. URL https://api.semanticscholar.org/CorpusID:168169753.

Kim, S., Chen, J., Cheng, T., Gindulyte, A., He, J., He, S., Li, Q., Shoemaker, B. A., Thiessen, P. A., Yu, B., Zaslavsky, L., Zhang, J., and Bolton, E. E. PubChem 2023 update. *Nucleic Acids Research*, 51(D1):D1373–D1380, 10 2022. ISSN 0305-1048. doi: 10.1093/nar/gkac956. URL https://doi.org/10.1093/nar/gkac956.

Ko, S. M., Lee, S., Jeong, D.-W., Kim, H., Lee, C., Yim, S., and Han, S. Multitask extension of geometrically aligned transfer encoder, 2024a.

Ko, S. M., Lee, S., Jeong, D.-W., Lim, W., and Han, S. Geometrically aligned transfer encoder for inductive transfer in regression tasks. In *The Twelfth International Conference on Learning Representations*, 2024b. URL https://openreview.net/forum?id=3z60EWfh1p.

Kumar, A., Sharma, K., and Dixit, A. R. A review on the mechanical and thermal properties of graphene and graphene-based polymer nanocomposites: understanding of modelling and md simulation. *Molecular Simulation*, 46:136 – 154, 2020. URL https://api.semanticscholar.org/CorpusID:208703597.

Li, H., Zhao, D., and Zeng, J. Kpgt: Knowledge-guided pre-training of graph transformer for molecular property prediction. *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, 2022. URL https://api.semanticscholar.org/CorpusID:249431724.

Li, S., Zhou, J., Xu, T., Dou, D., and Xiong, H. Geomgcl: Geometric graph contrastive learning for molecular property prediction. *ArXiv*, abs/2109.11730, 2021. URL https://api.semanticscholar.org/CorpusID:237635268.

Lu, C., Liu, Q., Wang, C., Huang, Z., Lin, P., and He, L. Molecular property prediction: A multilevel quantum interactions modeling perspective. In *AAAI Conference on Artificial Intelligence*, 2019. URL https://api.semanticscholar.org/CorpusID:57927281.

Pan, S. J. and Yang, Q. A survey on transfer learning. *IEEE Transactions on Knowledge and Data Engineering*, 22:1345–1359, 2010. URL https://api.semanticscholar.org/CorpusID:740063.

Qian, C., Tang, H., Yang, Z.-J., Liang, H., and Liu, Y. Can large language models empower molecular property prediction? *ArXiv*, abs/2307.07443, 2023. URL https://api.semanticscholar.org/CorpusID:259924688.

Sadybekov, A. V. and Katritch, V. Computational approaches streamlining drug discovery. *Nature*, 616:673–685, 2023. URL https://api.semanticscholar.org/CorpusID:258336875.

Sushko, I., Novotarskyi, S., Körner, R., Pandey, A. K., Rupp, M., Teetz, W., Brandmaier, S., Abdelaziz, A., Prokopenko, V. V., Tanchuk, V. Y., et al. Online chemical modeling environment (ochem): web platform for data storage, model development and publishing of chemical information. *Journal of computer-aided molecular design*, 25:533–554, 2011.

Weininger, D. Smiles, a chemical language and information system. 1. introduction to methodology and encoding rules. *J. Chem. Inf. Comput. Sci.*, 28:31–36, 1988. URL https://api.semanticscholar.org/CorpusID:5445756.

Wieder, O., Kohlbacher, S. M., Kuenemann, M. A., Garon, A., Ducrot, P., Seidel, T., and Langer, T. A compact review of molecular property prediction with graph neural networks. *Drug discovery today. Technologies*, 37:1–12, 2020. URL https://api.semanticscholar.org/CorpusID:230528669.

Yao, X., Liang, S., Han, S., and Huang, H. Enhancing molecular property prediction via mixture of collaborative experts. *ArXiv*, abs/2312.03292, 2023. URL https://api.semanticscholar.org/CorpusID:265696764.

You, Y., Chen, T., Sui, Y., Chen, T., Wang, Z., and Shen, Y. Graph contrastive learning with augmentations. *ArXiv*, abs/2010.13902, 2020. URL https://api.semanticscholar.org/CorpusID:225076220.

Zhuang, F., Qi, Z., Duan, K., Xi, D., Zhu, Y., Zhu, H., Xiong, H., and He, Q. A comprehensive survey on transfer learning. *Proceedings of the IEEE*, 109:43–76, 2019. URL https://api.semanticscholar.org/CorpusID:207847753.

A. Dataset

In this section, we provide the task set used for 40 task molecular property prediction, with their full name and corresponding number of data points.

Table 3. Dataset configuration used for the experiment of 40 tasks property prediction.

Property	Abbreviation	data points
Heat of vaporization	HV	1504
Viscosity	VS	1307
Surface tension	ST	977
Density	DS	3079
Boiling point	BP	8044
Refractive index	RI	11143
Melting point	MP	22901
LogP	LP	28268
Abraham descriptor S	AS	1915
Dielectric constant	DK	999
Dipole moment	DM	11224
Flash point	FP	9409
Ionization potential	IP	898
pKa	PKA	9514
Polarizability	POL	457
Vapor pressure	VP	4262
Absorbance maximum wavelength	AW	11896
Critical temperature	CT	2414
Heat of combustion	HC	2118
Hydration free energy	HF	648
Lower flammability limit temperature	LF	1646
HOMO energy level	HM	97262
LUMO energy level	LM	97262
Molar heat capacity (liquid)	CPL	387
Molar heat capacity (gas)	CPG	264
Molar heat capacity (solid)	CPS	218
Molar volume (liquid)	MVL	8513
Molar volume (solid)	MVS	218
Heat of vaporization	HVC	1957
Critical pressure	CTP	3007
Critical volume	CTV	2413
Gibbs Energy of Formation for Ideal gas	GEF	1828
Magnetic Susceptibility	MAS	432
Net Standard State Enthalpy of Combustion	NEC	1182
Parachor	PAR	960
Radius of Gyration	ROG	1370
Solubility Parameter	SPA	1509
Standard State Absolute Entropy	SAE	1072
Standard State Enthalpy of Formation	SEF	1638
Upper Flammability Limit Temperature	UFT	1443