Systems & Control Transactions

Original Research Article
Peer Reviewed Conference Proceeding
ESCAPE 36 - European Symposium on Computer Aided Process Engineering
Sheffield, UK. 21-24 June 2026

PSE PRESS

Evaluating and adapting modelling strategies for data-driven prediction of solvent effects on reaction barriers

Daeun Shin[ab], Lingfeng Gui[c], Jonggeol Na[d], Won Bo Lee[a], and Lauren Ye Seol Lee[b]*

[a] Department of Chemical and Biological Engineering, Seoul National University, Seoul 08826, Republic of Korea
[b] Department of Chemical Engineering, University College London, Torrington Place, London, WC1E 7JE, United Kingdom
[c] School of Engineering and Physical Sciences, Heriot-Watt University, Edinburgh, Scotland EH14 4AS, United Kingdom
[d] Department of Chemical Engineering and Materials Science, Ewha Womans University, Seoul 03760, Republic of Korea
* Corresponding Author: lauren.lee@ucl.ac.uk.

ABSTRACT

Predicting solvent effects on reaction activation barriers is central to understanding chemical reactivity and reaction kinetics, and guiding solvent selection. The solvent-induced change in activation free energy ($\Delta\Delta G^{\ddagger}_{\mathrm{solv}}$) provides a quantitative descriptor of this effect, but remains costly to evaluate across vast reaction-solvent spaces, using quantum mechanical methods. Recent data-driven models have enabled prediction of solvent effects. However, most typically rely on two-dimensional representation of reactions and do not explicitly encode sufficient reaction context, such as transition-state information, or three-dimensional structural changes along the reaction, resulting in limited generalizability and predictive accuracy. In this study, systematic evaluation is presented of modelling strategies for predicting $\Delta\Delta G^{\ddagger}_{\mathrm{solv}}$, with a focus on the role of reaction-state representation, input-geometry fidelity, and input modality. Using a large reaction-solvent dataset, models based on two-dimensional condensed reaction graphs are compared with models incorporating three-dimensional geometries of reactants, transition states, and products. The sensitivity of geometry-based models to structural accuracy is assessed by replacing quantum-chemically optimized transition states with structures predicted by a generative model. In addition, a dual-modality architecture combining two-dimensional graph-based and three-dimensional geometry-based representations is examined. The results show that explicit inclusion of both reactant and transition-state geometries leads to improved prediction accuracy relative to representations based on reaction endpoints or transition states alone. However, model performance depends strongly on the fidelity of the input geometries, with substantial degradation observed when low-quality structures are used. The dual-modality approach partially mitigates this sensitivity by adaptively reweighting two-dimensional and three-dimensional information, leading to performance recovery under low-fidelity conditions.

Keywords: Solvent effect, Solvation free energy of reaction, Transition state, 3D geometry, Multi-modality

1. INTRODUCTION

Solvents play a central role in chemical reactions because preferential stabilization of transition state (TS) relative to the reactant (R) can shift reaction energy barriers and alter reaction rate and selectivity. This solvent effect can be quantified by the solvation contribution to the activation free energy, defined as $\Delta\Delta G^{\ddagger}_{\mathrm{solv}} = \Delta G^{\mathrm{TS}}_{\mathrm{solv}} - \Delta G^{\mathrm{R}}_{\mathrm{solv}}$. The solvation free energy of reaction, $\Delta\Delta G^{\ddagger}_{\mathrm{solv}}$ can serves as a key performance descriptor, making it possible to compare solvent effects across reaction-solvent combinations and therefore useful for solvent selection. Traditionally, $\Delta\Delta G^{\ddagger}_{\mathrm{solv}}$ has been calculated via quantum mechanics/molecular mechanics (QM/MM) simulations or implicit solvation models, including such as the Solvation Model based on Density (SMD) and the COnductor-like Screening Model for Real Solvent (COSMO-RS). While these methods can provide reliable estimates, their high computational cost often limits scalability to a large and diverse reaction-solvent space [1]. In particular,

locating and optimising TS structures represents a major bottleneck, making systematic evaluation across many reactions and solvents impractical [2, 3].

In recent years, data-driven approaches have emerged as efficient alternatives by enabling rapid estimation of solvent effects at substantially reduced computational cost. Early work in this area has largely focused on predicting solvation energy of individual molecules [4-6] and extension to reaction kinetics have been more limited. Existing approaches for modelling solvent effect on activation energies broadly fall into two classes. The first class consider a single fixed reaction, treating solvent as the sole variable [7, 8]. While this yields meaningful results within specific reaction classes, it faces a fundamental limitation in generalizability across diverse reaction-solvent combinations. The second class aims to model general reaction-solvent combinations by using both reaction and solvent as inputs, and hence covering a broader chemical space [9, 10].

Recent progress in the latter direction has been driven by the effectiveness of two-dimensional (2D) reaction representations for modelling reaction properties, such as activation energies and reaction yields [10-12]. These representations include condensed reaction graph (CGR) [13], which is obtained by superposing 2D molecular graphs of R and product (P), and reaction fingerprint, which is generated by open-source software RDKit [14, 15] or deep learning models [16, 17]. Such representations have been used as input and achieved high predictive accuracy by effectively encoding structural changes resulting from reactions. This early success presents a promising modelling strategy of solvent effect on activation energy by combining 2D reaction representations with molecular representations of solvent. For example, Chung and Green [9] developed a data-driven model that can predict solvent-induced changes in activation energy utilizing CGR and 2D molecular graph of solvent. By integrating reaction and solvent embeddings independently encoded by directed message passing neural network (D-MPNN) encoders, the model achieved state-of-the-art performance in both $\Delta\Delta G^{\ddagger}_{\mathrm{solv}}$ calculated by COSMO-RS and experimental relative rate constant (k_{rel}) prediction. However, these representations primarily capture topological differences between R and P and do not explicitly encode information along the reaction coordinate or structural features of the TS. Moreover, 2D representations cannot uniquely determine three-dimensional (3D) properties, such as charge redistribution and spatial arrangement of functional groups, which are known to influence solvation and activation energetics. As a result, current models may struggle to generalize across reaction types and capture key physical factors governing solvent effects on reaction barriers. While employing TS geometries as input can provide richer information of reaction context, this has been considered impractical due to the difficulties in computation, as we described. To address this limitation, generative models that predict TS geometries have recently been actively proposed as a practical alternative [18-21]. These models generate TS geometries from relatively accessible inputs, e.g. CGR and geometries of R and P, at a fraction of the computational cost of QM methods. However, to the best of our knowledge, the effectiveness of explicit incorporation of the low-cost generated TS geometries for modelling solvent effect on activation energy has yet to be explored.

In this study, we develop systematic modelling strategies for $\Delta\Delta G^{\ddagger}_{\mathrm{solv}}$ prediction that explicitly address these limitations. Rather than proposing a new architecture, we analyse how predictive performance depends on three controlled aspects of model input design: (i) the reaction-state(s) used to represent the reaction (R, TS, and/or P), (ii) the fidelity of the associated 3D geometries, and (iii) the representation modality, including 2D graph-based, 3D geometry-based, and combined dual-modality inputs. Using a large reaction-solvent dataset derived from COSMO-RS calculations, we compare $\Delta\Delta G^{\ddagger}_{\mathrm{solv}}$prediction performance on unseen reactions of a conventional 2D model using CGR with models that explicitly incorporate geometries of various input configurations for reactants, transition states, and products. By replacing quantum-chemically optimised structures with geometries locally optimised by force field or predicted from a generative model, the sensitivity of the model to geometry quality is assessed. Dual-modality approaches combining 2D graph-based and 3D geometry-based representations are implemented to further explore the robustness of the models depending on fidelity of input geometry.

2 METHOD

2.1 Molecular and reaction representation

To investigate how input design influences model performance, we varied three factors systematically: (i) the state(s) of the reaction path encoded as input (R, TS, and/or P), (ii) the fidelity of the associated 3D geometries, and (iii) the modality of the representation i.e., whether the model used 2D graph-based inputs, 3D geometry-based inputs, or both.

For 2D representations, reactions were encoded as CGR [13], which represent the transformation between reactants and products within a single graph. CGRs were generated from atom-mapped SMILES using Chemprop [22]. Solvents were represented as standard 2D molecular graphs. Initial atom and bond features for both CGRs and solvent graphs were constructed using standard RDKit descriptors, including atomic number, bond order, formal charge, and aromaticity.

To incorporate explicit structural information along the reaction coordinate, 3D molecular geometries were

used to construct reaction representations based on different states of the reaction path. Three reaction-path configurations were considered (Figure 1**a**): (1) an R-P representation, which encodes the reaction using the 3D geometries of reactants (R) and products (P); (2) a TS-only representation, which encodes the reaction using only the 3D geometry of the transition state (TS); and (3) R-TS representation, which encodes the reaction using the 3D geometries of both reactants (R) and transition states (TS). These configurations enable a controlled comparison of how explicit inclusion of TS information and reaction-path context affects model behaviour.

Input geometries for 3D-based representations were further grouped by fidelity to evaluate sensitivity to structural accuracy. *High-fidelity* geometries used R, P, and TS structures optimised in the gas phase at the ωB97X-D3/def2-TZVP level of theory from the Grambow dataset [23]. *Moderate-fidelity* geometries used R and P structures from the Grambow dataset, while TS geometries were predicted using React-OT from the corresponding high-fidelity R and P structures. *Low-fidelity* geometries used R and P structures generated using OpenBabel [24] and locally optimised with the generalized Amber force field (GAFF), and TS geometries predicted by React-OT from these lower-cost R and P structures. Note that React-OT is a flow-matching-based generative model that predicts TS structures from interpolated R and P geometries [19] and was selected due to its deterministic predictions and computational efficiency. For solvents, 3D structures generated using the OpenBabel force field were used consistently across all fidelitiles.

2.2 Dataset for model training and evaluation

We used the $\Delta\Delta G^{\ddagger}_{\mathrm{solv}}$ dataset proposed by Chung and Green [9]. In that work, COSMO-RS calculations were performed using optimised gas-phase geometries of R and TS from the Grambow dataset, spanning 500, 000 reaction-solvent combinations with 13, 227 unique reactions and 295 solvents. The dataset both contains neutral closed-shell and free radical reactions, restricted to systems composed of H, C, O, and N with at most seven heavy atoms (11, 876 closed-shell; 1, 351 radical). To evaluate generalization to unseen reactions, 5% of the reaction were randomly selected and reserved as a test set. The remaining data were randomly split into a 90% training set and a 10% validation set.

2.3 Model architecture

Based on the molecular and reaction representations defined, three model architectures, namely, a 2D graph-based baseline model, a 3D geometry-based equivariant model, and a dual-modality model combining both representations. All architectures were adopted from existing frameworks and adapted to ensure a consistent comparison across input configurations.

2D-graph-based baseline model: As a baseline, we adopted a directed message passing neural network (D-MPNN) architecture [25] previously used for reaction-solvent modelling. This model independently encodes for the reaction and solvent by passing the 2D CGR and 2D solvent graph through separate D-MPNNs. A four-layer D-MPNN was used for the CGR to capture changes in atom and bond features associated with reaction, while a two-layer D-MPNN was used for the solvent graph owing to its comparatively simple structure. The resulting reaction and solvent embeddings were concatenated and passed to a feed-forward predictor consisting of six fully connected layers to generate the final prediction.

Geometry-based equivariant model: We adopted TorchMD-Net [26] as an SE(3)-equivariant encoder. TorchMD-Net takes atomic numbers and Cartesian coordinates as inputs and constructs representations using attention mechanisms based on interatomic distance and relative position vectors, without relying on explicit bond connectivity. An eight-layer equivariant encoder was applied to all molecular species. The representations obtained from the encoder for R, TS, and P were concatenated and passed through two fully connected layers to form reaction-level embedding. This embedding was then concatenated with the solvent representation and passed through the same six-layer feed-forward predictor used in the baseline model.

Duality-modality model: To evaluate whether combining 2D and 3D representations improves robustness [27, 28], we implemented a dual-modality architecture that integrates the D-MPNN-based 2D encoders and the TorchMD-Net-based SE(3)-equivariant 3D encoder. In this setting, representations obtained independently from the 2D and 3D streams were combined using a late-fusion strategy. Learnable scalar weights were applied to each modality stream, allowing the model to adjust their relative contributions during training. Specifically, two scalar parameters, i.e., w_{2D} and w_{3D}, initialized to zero, were used. These weights are passed through a softmax function to normalise them and multiplied by the embeddings from each modality stream to obtain a fused representation. The fused representation was then passed to the same feed-forward predictor used in the single-modality models (Figure 4**a**).

All models were trained for 80 epochs using a batch size of 30. Optimisation was performed with the Adam optimiser and Noam learning-rate scheduler, with an initial learning rate of 1.8×10^{4}. Mean squared error (MSE) was used as the loss function. Model implementation was performed in PyTorch 2.4.0, and training was conducted using PyTorch Lightning 1.8.6.

3. RESULTS

3.1 Effect of including TS geometries

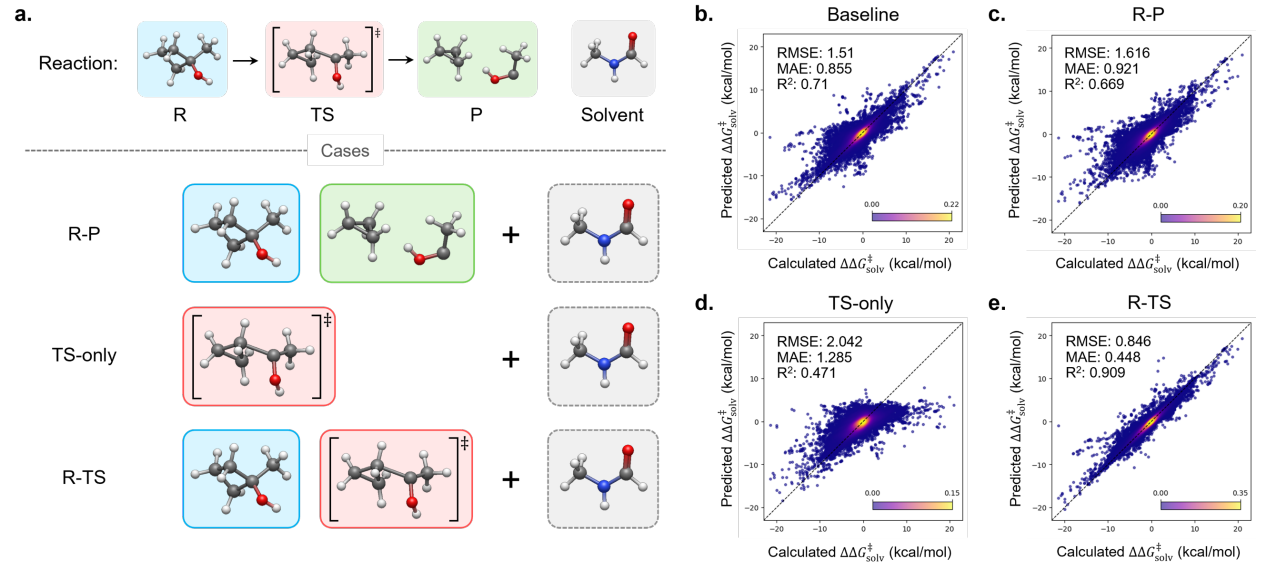

Figure 1. Overview of input configurations and prediction performance for different representations. (a) Schematic of the three 3D input configurations considered in this study: R-P, which uses the geometries of R and P; TS-only, which uses only the TS geometry; and R–TS, which uses the geometries of both R and TS. In all cases, the solvent is represented separately and combined with the reaction representation. (b-e) Parity plots comparing predicted and calculated $\Delta\Delta G^{\ddagger}_{\mathrm{solv}}$ values for (b) the 2D graph-based baseline model, (c) the R-P model, (d) the TS-only model, and (e) the R–TS model. Root mean square error (RMSE), mean absolute error (MAE), and coefficient of determination (R^2) are reported for each model.

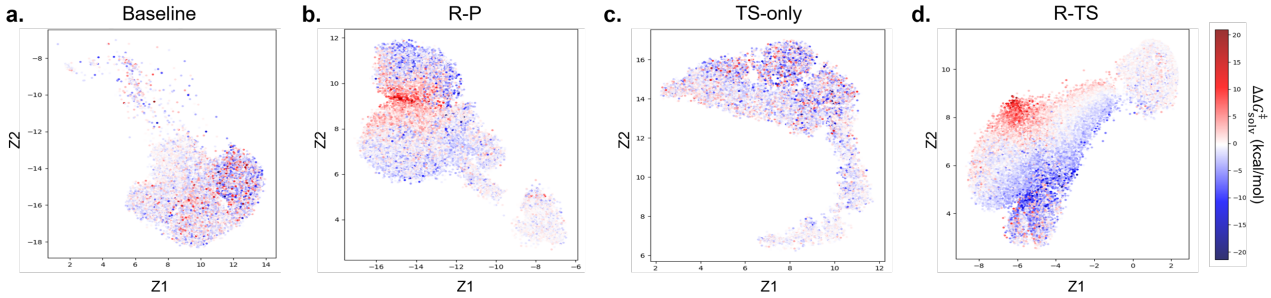

Figure 2. UMAP visualisations of the learned reaction-solvent representations for different input configurations. Latent embeddings are shown for (a) the 2D graph-based baseline model and the 3D models using (b) R-P, (c) TS-only, and (d) R-TS. Points are coloured by the corresponding $\Delta\Delta G^{\ddagger}_{\mathrm{solv}}$ values.

To evaluate the effect of including TS geometries on $\Delta\Delta G^{\ddagger}_{\mathrm{solv}}$ prediction, we compared the model performance using three distinct combinations of R, TS, and P geometries defined in Section 2.1—R-P, TS-only, and R-TS, with the 2D-based baseline model. The input configurations are summarised in Figure 1**a**, and all models in this comparison used high-fidelity geometries from the Grambow dataset.

Notably, the model using 3D geometries of R and P (R-P) exhibited lower prediction accuracy than the 2D-based baseline (Figure 1**b, c**). This result indicates that providing endpoint geometries alone does not improve prediction of solvent-induced activation free energy shifts. Although 3D molecular structures contain detailed geometric information, the absence of explicit TS information limits their relevance for describing kinetically meaningful structural changes. As a result, replacing 2D CGR inputs with 3D R–P geometries increase input complexities without improving representation of features directly associated with the target quantity.

The model using only TS geometry (TS-only) showed significantly lower prediction performance compared to the baseline, even though TS is a critical determinant of reaction kinetics (Figure 1**d**). This result likely stems from the fact that understanding reaction process requires not only TS geometry but also R structure as a reference state of change. The omission of the information of the reference state has led to overfitting, confirming that TS structure alone is insufficient to capture the relative structural evolution that is necessary to

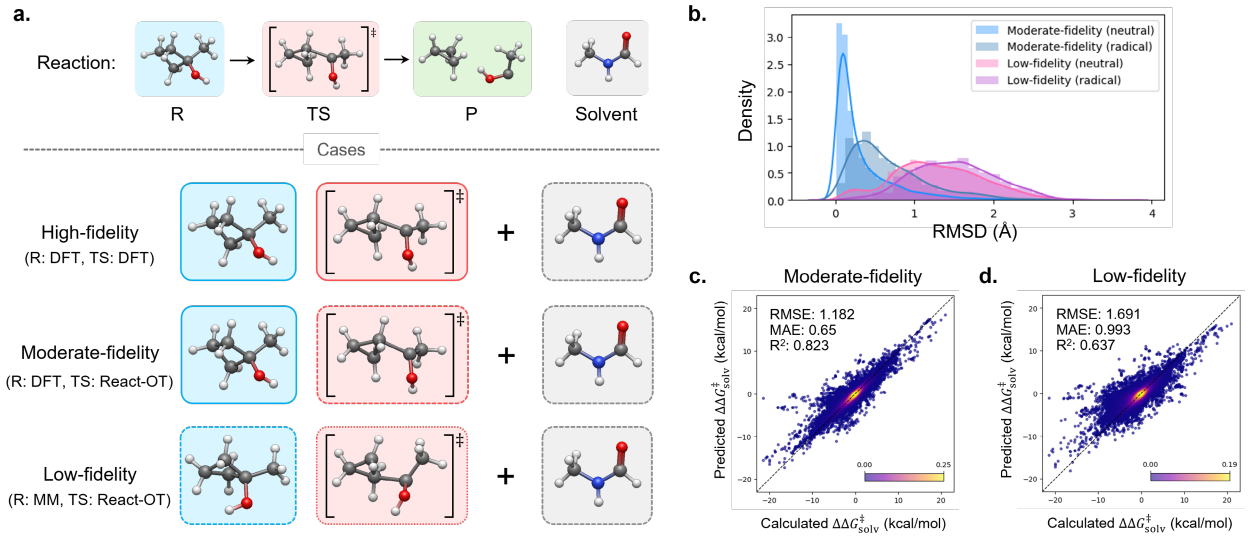

Figure 3. Effect of input structure fidelity on TS geometry accuracy and prediction performance. (a) Schematic illustration of the three input-geometry fidelity levels considered for the R–TS configuration. (b) Distributions of TS root-mean-square deviations (RMSDs) relative to the DFT reference structures for moderate- and low-fidelity cases, shown separately for neutral and radical reactions. (c,d) Parity plots comparing predicted and calculated

define a reaction path.

In contrast, incorporating both R and TS 3D geometries (R-TS) achieved the highest prediction accuracy (Figure 1**e**). This indicates that the model implicitly captured path dependent features, such as geometric distortion and charge redistribution, from the structural data of R and TS. Thus, explicitly incorporating R and TS 3D structures enables the model to encode structural differences between R and TS, which are directly relevant to solvent-induced changes.

To further examine how different input configurations influence the learned representation, the final reaction-solvent embeddings were visualised using Uniform Manifold Approximation and Projection (UMAP) [29]. For the baseline model and the TS-only model, the latent representations were broadly scattered and showed no clear patterns with respect to the target values (Figure2**a, c**). The R-P model showed a locally clustered region for reaction-solvent combinations with high $\Delta\Delta G^{\ddagger}_{\mathrm{solv}}$, but remained disordered for moderate to small values (Figure 2**b**), consistent with its larger prediction errors observed in this region. In contrast, the latent space learned by the R-TS model exhibited the clearest global ordering with respect to $\Delta\Delta G^{\ddagger}_{\mathrm{solv}}$, characterised by a continuous gradient across the embedding (Figure 2**d**). This suggest that the combined use of R and TS geometries yields representations more strongly associated with solvent-induced activation free energy shifts.

3.2 Effect of input structure fidelity

In previous section, we showed that incorporating both R and TS geometries improves prediction performance when high-fidelity structures are available. To assess the robustness of this approach under more practical conditions, we next evaluated how prediction performance depends on the fidelity of the input geometries, using the three structure-fidelity levels defined in Section 2.1.

As summarised in Figure 3**a**, model performance was compared across high-, moderate-, and low-fidelity input geometries. In the high-fidelity case, both R and TS geometries were taken from DFT-optimised structures in the Grambow dataset. In the moderate- and low-fidelity cases, TS geometries were generated using React-OT from progressively lower-fidelity R and P structures, as described in Section 2.1. Consistent with this construction, the TS RMSD relative to the DFT reference structures was smaller in the moderate-fidelity case and substantially larger in the low-fidelity case (Figure 3**b**). The broader RMSD distributions observed for radical reactions further indicate increased difficulty in accurately predicting TS geometries for open-shell systems.

Prediction accuracy decreased systematically as input structure fidelity was reduced (Figure 3**c, d**). In the low-fidelity case, the model performed worse than the 2D-based baseline, indicating that TS geometries generated from inaccurate R and P structures do not provide reliable information for learning solvent-induced changes in the activation free energy. In contrast, the moderate-fidelity case maintained superior performance relative to the baseline, demonstrating that predicted TS geometries can still be informative when the underlying R and P structures preserve essential geometric features.

Overall, these results show that models explicitly

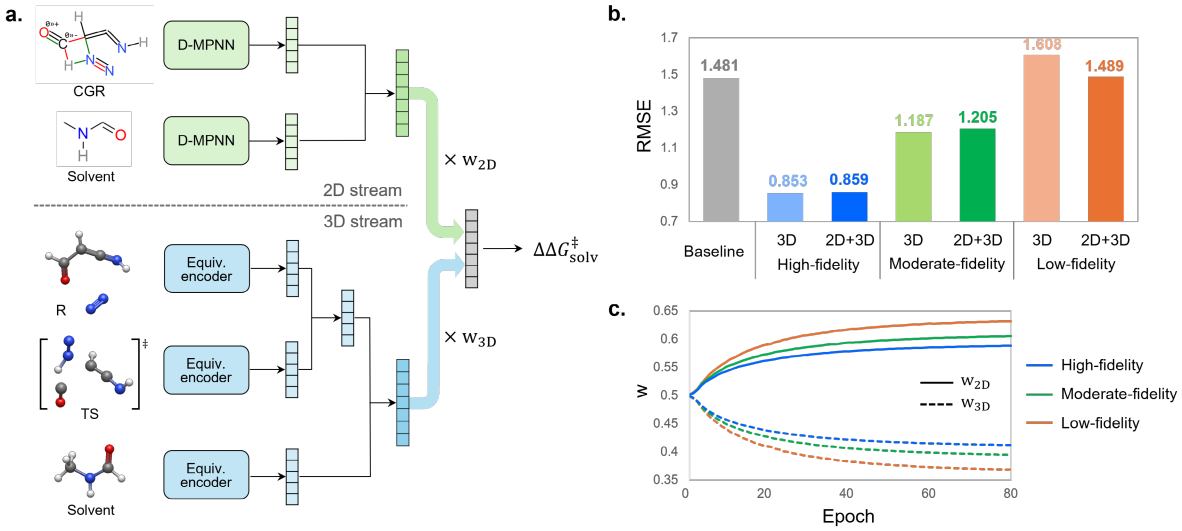

Figure 4. Dual-modality model architecture and performance across input-geometry fidelity levels. (a) Schematic illustration of the dual-modality architecture integrating 2D graph-based representations and 3D geometry-based representations. (b) Prediction performance, measured by RMSE, for the 2D-based baseline model, the 3D-only model, and the dual-modality (2D + 3D) model across high-, moderate-, and low-fidelity input geometries. (c) Evolution of the learned modality weights for the 2D and 3D streams during training for each fidelity level.

incorporating R and TS geometries are sensitive to the fidelity of the input structures. Nevertheless, the performance observed in the moderate-fidelity case indicates that practical modelling strategies based on low-cost TS generation are feasible, provided that the predicted structures retain key geometric characteristics of the activated complex.

3.3 Effect of input modality

Because prediction performance of 3D-based models depends strongly on input-geometry fidelity, we next evaluated whether combining 2D and 3D representations can improve robustness to reduced structural accuracy. Specifically, we examined a dual-modality approach that integrates 2D graph-based representations with 3D geometry-based representations, as defined in Section 2.3.

The dual-modality model integrates 2D graph-based and 3D geometry-based representations using the late-fusion architecture described in Section 2.3 (Figure 4**a**). Model performance was evaluated for high-, moderate-, and low-fidelity input geometries and compared against the corresponding 3D-only and 2D-based baseline models.

For the high- and moderate-fidelity cases, the dual modality model showed only marginal difference relative to the 3D-only model as can be seen in Figure 4**b**. This suggests that when reliable 3D structures are available, the 3D representation alone captures sufficient information for $\Delta\Delta G^{\ddagger}_{\mathrm{solv}}$, limiting the additional contribution of 2D graph-based inputs. In contrast, in the low-fidelity case, the dual modality model recovered prediction performance to a level comparable to the 2D-based baseline and outperformed the 3D-based model (Figure 4**b**). Analysis of the learned modality weights (w) provides further insight into this behaviour (Figure 4**c**). In the low-fidelity case, the weight assigned to the 2D modality increased progressively during training while the 3D weight decreased, indicating a shift toward graph-based representations as the reliability of 3D geometries decreased. Conversely, in the high-fidelity case, the modality weights stabilised early and the gap between w_{2D} and w_{3D} remained smallest, indicating a more balanced contribution of the two modalities.

Overall, these results indicate that dual-modality models can improve robustness to reduced input-geometry fidelity by adaptively reweighting information from different representations. The observation that performance in the low-fidelity case does not exceed that of the 2D-based baseline highlights a limitation of the current late-fusion strategy, suggesting that more effective integration schemes may be required to achieve synergistic gains from combined 2D and 3D inputs.

4. CONCLUSIONS

We presented systematic modelling strategies for solvent-induced change in reaction activation energy, specifically for $\Delta\Delta G^{\ddagger}_{\mathrm{solv}}$ prediction. Comprehensive analysis of prediction based on thee controlled aspects of model input design was performed: (i) the reaction-state(s) used to represent the reaction, (ii) the fidelity of the associated input geometries, and (iii) the

representation modality of 2D, 3D, and combined. The significant improvement in predictive accuracy was observed through explicit incorporation of the geometries of both R and TS. Furthermore, the visualised latent space learned by the R-TS model exhibited the clearest trend with respect to $\Delta\Delta G^{\ddagger}_{solv}$. This proves that the structural differences encoded from the R and TS geometries yield representations strongly relevant to solvent effect on activation free energy. The performance of the 3D geometry-based model was strongly influenced by input geometry fidelity. When low-fidelity geometries were used, the model showed degraded performance compared to the 2D graph-based baseline model. However, the performance under the moderate-fidelity case maintained superior performance relative to the baseline. This indicates the feasibility of the practical modelling strategies based on predicted TS geometries and concurrently highlights the importance of preserving key geometric characteristics of the activated complex. Finally, we implemented a dual modality approach that integrates 2D graph-based and 3D geometry-based representations to investigate the robustness of the model to reduced fidelity of input geometries. Under the low-fidelity cases, the model exhibited a notable recovery of prediction performance to a level comparable to the 2D-based baseline and outperformed the 3D-based model. Further analysis provided valuable insight that the improved robustness resulted from the adaptive reweighting of the information from different modalities based on their reliability. Nevertheless, the current simple late-fusion strategy failed to facilitate synergistic interaction between modalities, limiting the performance in the low-fidelity case to a similar level to the 2D-based baseline. This highlights that future works should focus on developing multi-modal frameworks that enable complementary integration of information using sophisticated techniques, such as contrastive learning and cross-attention mechanisms. We believe that such advancement can provide a practical strategy of modelling solvent-induced changes in reaction barriers by harnessing the benefits of 3D geometries while enhancing the robustness from noise-resistant inputs.

ACKNOWLEDGEMENTS

This research was supported by Korea Institute for Advancement of Technology (KIAT) grant funded by the Ministry of Trade, Industry & Energy (MOTIE), Korea Government (RS-2024-00436106, Human Resource Development Program for Industrial Innovation).

AUTHOR IDENTIFIERS

Author ORCIDs:
Shin D: 0009-0000-9044-8113
Gui L: 0000-0003-1957-1957
Na J: 0000-0002-1106-9500
Lee WB: 0000-0001-7801-083X
Lee YS: 0000-0003-1847-7838

REFERENCES

1. Spiekermann KA, Pattanaik L, Green WH. Fast predictions of reaction barrier heights: toward coupled-cluster accuracy. J. Phys. Chem. A 126:3976-3986 (2022). https://doi.org/10.1021/acs.jpca.2c02614
2. Zhao Q, Hsu HH, Savoie BM. Conformational sampling for transition state searches on a computational budget. J. Chem. Theory Comput. 18:3006-3016 (2022). https://doi.org/10.1021/acs.jctc.2c00081
3. Jackson R, Zhang W, Pearson J. Tsnet: predicting transition state structures with tensor field networks and transfer learning. Chem. Sci. 12:10022-10040 (2021). https://doi.org/10.1039/d1sc01206a
4. Ferraz-Caetano J, Teixeira F, Cordeiro MNDS. Explainable supervised machine learning model to predict solvation gibbs energy. J. Chem. Inf. Model. 64:2250-2262 (2023). https://doi.org/10.1021/acs.jcim.3c00544
5. Chung, Y., et al., Group contribution and machine learning approaches to predict Abraham solute parameters, solvation free energy, and solvation enthalpy. Journal of Chemical Information and Modeling, 2022. **62**(3): p. 433-446.
6. Low K, Coote ML, Izgorodina EI. Explainable solvation free energy prediction combining graph neural networks with chemical intuition. J. Chem. Inf. Model. 62:5457-5470 (2022). https://doi.org/10.1021/acs.jcim.2c01013
7. Gui L, Yu Y, Oliyide TO, Siougkrou E, Armstrong A, Galindo A, Sayyed FB, Kolis SP, Adjiman CS. Integrating model-based design of experiments and computer-aided solvent design. Computers & Chemical Engineering 177:108345 (2023). https://doi.org/10.1016/j.compchemeng.2023.108345
8. Struebing H, Ganase Z, Karamertzanis PG, Siougkrou E, Haycock P, Piccione PM, Armstrong A, Galindo A, Adjiman CS. Computer-aided molecular design of solvents for accelerated reaction kinetics. Nature Chem 5:952-957 (2013). https://doi.org/10.1038/nchem.1755
9. Chung Y, Green WH. Machine learning from quantum chemistry to predict experimental solvent effects on reaction rates. Chem. Sci. 15:2410-2424 (2024). https://doi.org/10.1039/d3sc05353a
10. Jorner K, Brinck T, Norrby PO, Buttar D. Machine

learning meets mechanistic modelling for accurate prediction of experimental activation energies. Chem. Sci. 12:1163-1175 (2021). https://doi.org/10.1039/d0sc04896h
11. Schwaller P, Vaucher AC, Laino T, Reymond JL. Prediction of chemical reaction yields using deep learning. Mach. Learn.: Sci. Technol. 2:015016 (2021). https://doi.org/10.1088/2632-2153/abc81d
12. Probst D, Schwaller P, Reymond JL. Reaction classification and yield prediction using the differential reaction fingerprint DRFP. Digital Discovery 1:91-97 (2022). https://doi.org/10.1039/d1dd00006c
13. Heid E, Green WH. Machine learning of reaction properties via learned representations of the condensed graph of reaction. J. Chem. Inf. Model. 62:2101-2110 (2021). https://doi.org/10.1021/acs.jcim.1c00975
14. Dobbelaere MR, Lengyel I, Stevens CV, Van Geem KM. Rxn-insight: fast chemical reaction analysis using bond-electron matrices. J Cheminform 16: (2024). https://doi.org/10.1186/s13321-024-00834-z
15. van Gerwen P, Briling KR, Calvino Alonso Y, Franke M, Corminboeuf C. Benchmarking machine-readable vectors of chemical reactions on computed activation barriers. Digital Discovery 3:932-943 (2024). https://doi.org/10.1039/d3dd00175j
16. Schwaller P, Probst D, Vaucher AC, Nair VH, Kreutter D, Laino T, Reymond JL. Mapping the space of chemical reactions using attention-based neural networks. Nat Mach Intell 3:144-152 (2021). https://doi.org/10.1038/s42256-020-00284-w
17. Mswahili ME, Jeong YS. Transformer-based models for chemical SMILES representation: a comprehensive literature review. Heliyon 10:e39038 (2024). https://doi.org/10.1016/j.heliyon.2024.e39038
18. Kim S, Woo J, Kim WY. Diffusion-based generative AI for exploring transition states from 2D molecular graphs. Nat Commun 15: (2024). https://doi.org/10.1038/s41467-023-44629-6
19. Duan C, Liu GH, Du Y, Chen T, Zhao Q, Jia H, Gomes CP, Theodorou EA, Kulik HJ. Optimal transport for generating transition states in chemical reactions. Nat Mach Intell 7:615-626 (2025). https://doi.org/10.1038/s42256-025-01010-0
20. Duan C, Du Y, Jia H, Kulik HJ. Accurate transition state generation with an object-aware equivariant elementary reaction diffusion model. Nat Comput Sci 3:1045-1055 (2023). https://doi.org/10.1038/s43588-023-00563-7
21. Choi S. Prediction of transition state structures of gas-phase chemical reactions via machine learning. Nat Commun 14: (2023). https://doi.org/10.1038/s41467-023-36823-3
22. Heid, E., et al., *Chemprop: a machine learning package for chemical property prediction.* Journal of Chemical Information and Modeling, 2023. **64**(1): p. 9-17.
23. Grambow, C.A., L. Pattanaik, and W.H. Green, Reactants, products, and transition states of elementary chemical reactions based on quantum chemistry. Scientific data, 2020. **7**(1): p. 137.
24. O'Boyle NM, Banck M, James CA, Morley C, Vandermeersch T, Hutchison GR. Open babel: an open chemical toolbox. J Cheminform 3: (2011). https://doi.org/10.1186/1758-2946-3-33
25. Yang K, Swanson K, Jin W, Coley C, Eiden P, Gao H, Guzman-Perez A, Hopper T, Kelley B, Mathea M, Palmer A, Settels V, Jaakkola T, Jensen K, Barzilay R. Analyzing learned molecular representations for property prediction. J. Chem. Inf. Model. 59:3370-3388 (2019). https://doi.org/10.1021/acs.jcim.9b00237
26. Thölke, P. and G. De Fabritiis, Torchmd-net: equivariant transformers for neural network based molecular potentials. arXiv preprint arXiv:2202.02541, 2022.
27. Luo, Y., et al., *Molfm: A multimodal molecular foundation model.* arXiv preprint arXiv:2307.09484, 2023.
28. Yu, Q., et al., *Multimodal molecular pretraining via modality blending.* arXiv preprint arXiv:2307.06235, 2023.
29. McInnes, L., J. Healy, and J. Melville, *UMAP: uniform manifold approximation and projection for dimension reduction. arXiv.* arXiv preprint arXiv:1802.03426, 2018. **10**.