# From Reverse Phase Chromatography to HILIC: Graph Transformers Power Method-Independent Machine Learning of Retention Times

Cailum M. K. Stienstra,[a] Emir Nazdrajić,[a] and W. Scott Hopkins[a,b,c*]

[a] *Department of Chemistry, University of Waterloo, Waterloo, Ontario, N2L 3G1, Canada,*

[b]*WaterFEL Free Electron Laser Laboratory, University of Waterloo, Waterloo, Ontario, N2L 3G1, Canada*

[c]*Centre for Eye and Vision Research, Hong Kong Science Park, New Territories, 999077, Hong Kong.*

**Corresponding Author*

**ABSTRACT:** Liquid chromatography (LC) is a cornerstone of analytical separations, but comparing the retention times (RTs) for different LC methods is difficult because of variations in experimental parameters such as column type and solvent gradient. Nevertheless, RTs are powerful metrics in tandem mass spectrometry ($MS^2$) that can reduce false positive rates for metabolite annotation, differentiate isobaric species, and improve peptide identification. Here, we present Graphormer-RT, a novel graph transformer that performs the first "method-independent" prediction of RTs. We use the RepoRT dataset, containing 142,688 reverse phase (RP) RTs (191 methods) and 4,373 HILIC RTs (49 methods). Our best RP model achieved a test set mean average error (MAE) of 29.3±0.6 s, a significant improvement over the previous record (1 method). Our best performing HILIC model achieved a test MAE=42.4±2.9 s. Extending this proof-of-concept work could enable machine-optimization of automated LC workflows and *in silico* annotation of unknown analytes in LC-$MS^2$ measurements.

## Introduction

Liquid chromatography (LC) has long been a cornerstone of analytical separations, providing a reliable means to isolate metabolites, peptides, pharmaceuticals, and xenobiotics.[1–4] The retention time (RT) is the defining parameter of this technique, describing the amount of time required for a particular analyte to be carried by a solvent mixture (*i.e.*, a mobile phase) through a column packed with a stationary phase.[4] LC is commonly coupled to mass spectrometry (MS) to separate and characterize molecules in complex mixtures such as those analyzed in "-omics" studies (*i.e.*, proteomics, metabolomics, lipidomics, *etc.*).[4] In targeted MS experiments, predicted RTs have been used to differentiate isobaric lipids,[5] reduce false-positive annotation rate of small molecule $MS^2$ fragmentation spectra,[6] and increase the accuracy of peptide identification schema.[7,8] In general, LC-MS annotation using machine learning has been shown to meaningfully increases the confidence of small molecule,[9,10] metabolite,[10] peptide,[11,12] and cell/organelle interactome identities.[13]

Quantitative structure-property relationships (QSPRs) have been used for decades to derive relationships with and predict analyte RTs. However, despite the many machine learning (ML) models that have been reported to date, there exists no generalizable framework for the method-independent prediction of RT.[14–16] Comparison of chromatographic methods between systems is difficult because of variances in measured RTs due to differences in parameters such as column properties, mobile phase composition, gradient profile, flow rate, pH, temperatures, and matrix effects to name a few. Current, successful, deep learning frameworks utilize an internally consistent set of predictions for a singular chromatographic setup.[3,17] When the method is held constant, this multi-dimensional description of column and gradient is learned implicitly by the model because these parameters are invariant to all RTs.[3,17] This paradigm is embodied by the METLIN SMRT dataset,[3] a library of more than 80,000 small molecule RTs obtained using a single LC method. Employing this database, RTs have been predicted to within less than a minute using artificial neural networks (ANNs),[8] ensemble regressors,[8,18] Graph Neural Networks (GNNs),[3,17–22] and Convolutional Neural Networks (CNNs).[23,24] Prediction of RTs should be thought of as two complementary tasks: *(i)* learning the relationship between molecular structure and degree of interaction with the mobile and stationary phases and *(ii)* scaling those interactions to a given set of chromatographic conditions. Because the knowledge gained from *(i)* is consistent across chromatographic setups, transfer learning has been used to pass this information to setups where *(ii)* is learned upon finetuning.[23,25] This approach has been successfully applied for RTs in smaller datasets where the chromatographic conditions are internally consistent and the "method rescaling" is learned implicitly.[23,25] However, in these cases, the prediction error can be as much as double the first dataset (*ca*. 1-2 minutes) and questions about generalizability remain unanswered.[15,17]

Although progress in RT prediction is encouraging, there yet exists no method-independent RT prediction tool. To create an accurate method-independent model, one needs a sufficiently large RT prediction dataset that contains the necessary standardized method data (*e.g.*, intrinsic column data, gradient, *etc.*) and a model that can combine tasks (*i*) and *(ii)* in context of each other to produce sufficiently descriptive embeddings. In 2024, the RepoRT dataset was published,[16] providing the necessary chromatographic library and metadata needed to describe the total set of column conditions.

In this work we investigate the implementation of Graphormer to create the first method-independent predictive model of RTs for reverse phase (RP) and hydrophobic interaction liquid chromatography (HILIC).[26] Graphormer was developed by Ying *et al.* while working at Microsoft in

https://doi.org/10.26434/chemrxiv-2024-gh05j **ORCID:** https://orcid.org/0000-0001-7365-1558

2021, and it extends the transformer architecture to graph neural networks.[27] This architecture won the 2021 Open Catalyst challenge by utilizing a *global receptive field* and attention mechanisms that allow for highly contextual descriptions. [26,28,29] Graphormer has previously achieved state-of-the-art predictions across a wide variety of (bio)cheminformatics tasks.[28,30,31] We utilize empirical descriptions of chromatographic metadata from RepoRT that are passed to a flexible "pre-graph" chromatography encoder that derives a dense, learned representation of method data. This information is stored in a global graph node that allows method information to be considered *in context* of the molecular structure. We explore the strengths of Graphormer-RT in predictions for different methods (*i.e.*, different columns, gradients, manufacturers, *etc.*), and further assess model generalizability by excluding methods from the training library and testing model performance against the held-out methods. We also show that these formalisms can be extended to HILIC RTs. To the best of our knowledge, this work achieves the first method-independent RT predictions, as well as the first prediction of HILIC RTs for small molecules.

## Results and Discussion

### Model Performance

**RP Methods.** Models were trained using five-fold cross validation ($c = 5$) to optimize model configuration. Figure 1 shows a summary of the chromatographic metadata used in encoding, where encoding formalisms are discussed in more detail in the *Methods* section. The best performing Graphormer-RT model (see Figure 2) achieved a test mean average error (MAE) of 29.3 ± 0.6 s ($c = 5$). Although highly imperfect due to differences in number of retention times, methods, *etc.*, we compare our results to models trained on only the METLIN SMRT dataset as a coarse benchmark for our "method-independent" results. **Table 1** provides a comparison between several published models trained on the METLIN SMRT dataset with their associated cross-validated MAEs. If we compare our results to the best performing model (*i.e.*, the MPNN work by Osipenko *et al.*), which reported a MAE = 32.1 ± 0.6 s ($c$=5),[17] Graphormer-RT achieves predictions that are 9.5% or 4.7σ improved, well beyond the 99.9% confidence interval threshold. This substantial improvement is achieved while also making similarly accurate predictions for 190 other LC methods that are not considered by previous works. If we examine our model's test performance on only the SMRT dataset, we achieve test set error of MAE = 37.7 s ($n = 7{,}814$), which suggests that the improved generalization of Graphormer-RT comes at the expense of prediction accuracy for the SMRT library.

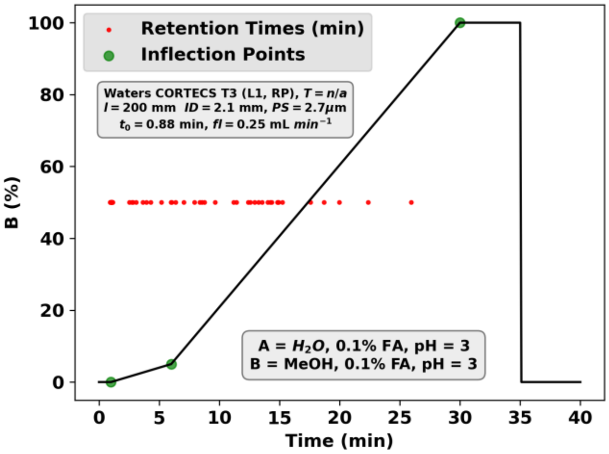

**Figure 1.** A summary of chromatographic method 0153 from RepoRT.[32] The column length ($l$), the column inner diameter ($ID$), the particle size ($PS$), the dead time ($t_0$), and the flow rate ($fl$) are characteristic method parameters. RTs and inflection points are plotted atop the gradient in red and green, respectively.

**Table 1.** Comparison of Graphormer-RT performance versus other LC prediction models with associated number of RTs ($N_{RT}$) and number of methods ($N_{methods}$). The variable, c, denotes the number of cross validation folds used in evaluation. Method mean absolute errors (MAEs) are provided but should not be compared directly due to differences in dataset composition.

| Model/Dataset Description (HPLC Type) | $N_{RT}$ ($N_{Methods}$) | Test MAE (s) |
|---|---|---|
| Graphormer-RT (RP) | 142,688<br>(191) | **29.3 ± 0.6 ($c$ = 5)** |
| MPNN-METLIN (RP)[17] | 77,977 (1) | 32.1 ± 0.6 ($c$ = 5) |
| GCN-METLIN (RP)[3] | 80,038 (1) | 57 ($c$ = 1) |
| 1D-CNN-METLIN (RP)[23] | 77,983 (1) | 34.7 ± 1.2 ($c$ =10) |
| Graphormer-RT (HILIC) | 4,373 (49) | **42.4 ± 2.9 ($c$ = 5)** |

Figure 3 depicts ML correlation diagrams for RP RT prediction. The correlation plot for the best performing Graphormer-RT split is shown in panel i, and panels ii-vi provide color coding to indicate column manufacturer, column length, $t_0$, mobile phase B, and flow rate, respectively. Figure 3ii shows that in our RP dataset, the most common manufactures are Waters and Agilent, where prediction accuracy on Agilent systems (MAE = 37.7 s, $n$ = 4,313) is much poorer than on Waters systems (MAE = 19.8 s, $n$ = 7,821). Whether these differences are related to the hardware, methods, or the chemicals systems, is unknown. Figure 3v shows that Graphormer-RT is slightly better at predicting methods using a methanol (MeOH) organic phase (MAE = 28.9 s) rather than acetonitrile (ACN) (MAE = 26.6 s). Figure 3iv shows how methods with smaller $t_0$ typically have shorter RTs and are better predicted ($t_0 < 1.0$ min, $n$ = 5,789, MAE = 16.6 s) than those with large $t_0$ ($t_0 \geq 1.0$ min, $n$ = 8,478, MAE = 36.8 s). Figure 3vi shows a similar but inverted effect, where low-flow rate ($fl$) systems, which typically have longer retention times, exhibit poorer prediction accuracy ($fl < 0.4$ mL/min, $n$ = 9,604, MAE = 36.0 s) compared to high flow rate systems ($fl \geq 0.4$ mL/min, $n$ = 4,663, MAE = 13.3 s).

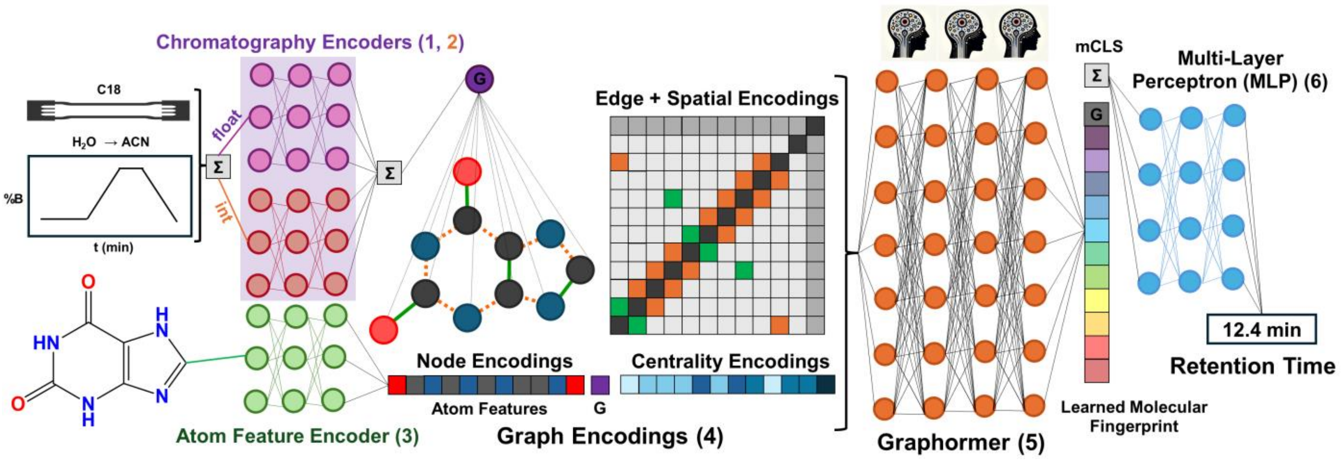

**Figure 2.** Schematic illustration of the Graphormer-RT architecture where the SMILES string for a molecule (xanthine, here) is mapped into a graph by the Atomic Feature Encoder (**3**) for node/atom embeddings and our combinatoric encoding scheme for edge encodings. Simultaneously, chromatographic metadata is passed into float (**1**) and integer (**2**) chromatography encoders, which generate learned, dense, representations that are added together and stored in the global chromatography node (**G**). Molecular graphs are then passed to Graphormer (**5**), whose centrality, spatial, and node encodings are illustrated alongside the model architecture. [26] The learned molecular representation generated by Graphormer is aggregated into the mCLS token and decoded via the output MLP (**6**), producing the final RT prediction.

These trends in flow rate and dead volume prediction error are likely influenced by the *heteroscedasticity* of MAE for retention time predictions. In other words, a 30 s absolute error is much more meaningful for a measured RT of 5 minutes compared to one at 25 minutes. Figure S1i plots MAE as a function of RT to better visualize this behaviour. In comparison, Figure S1ii plots the mean absolute percent error (MAPE) as a function of RT. Numerically, we observe that the MAE is a smaller fraction (5.1 %, see Figure S1) of the average test RT value ($RT_\mu$ = 565.0 s), compared to the calculated test MAPE (7.0 %), suggesting that MAE underestimates the relative contribution of smaller RTs. It is clear that MAPE is a more consistent homoscedastic error metric, which we believe provides a more appropriate measure to describe the per-RT prediction error independently of the gradient duration. However, we will report MAE here as well since this is the error metric commonly used in the literature.[3,17] Finally, to investigate consistency among RP data sets, we plotted the MAE for each individual RP method against the logarithm of the number of instances for that method in the test set (see Figure S2). Doing so revealed that the number of RTs for a given method that appear in the test set (and by extension, the entire dataset) does not bias the average performance of the model.

**External Method Validation.** To further investigate the generalization of Graphormer-RT, we predicted RTs for LC methods had not been used to train the model. This approach provides information regarding the extent to which the model is "memorizing" the LC method scaling factors from the training set (task *ii*; *vide supra*). The results for "held-out" external LC methods are reported in **Table 2**. Additional details are available in Supplementary Figures S3, S5-S9. For methods 0127 and 0275,[33,34] Graphormer-RT generalizes well, achieving MAEs of 43.9 ± 3.3 s and 42.7 ± 6.2 s ($c$ = *5*), respectively. While slightly worse than the accuracy of the global test set that was randomly split from the training set LC methods, the sub-minute accuracies inspire

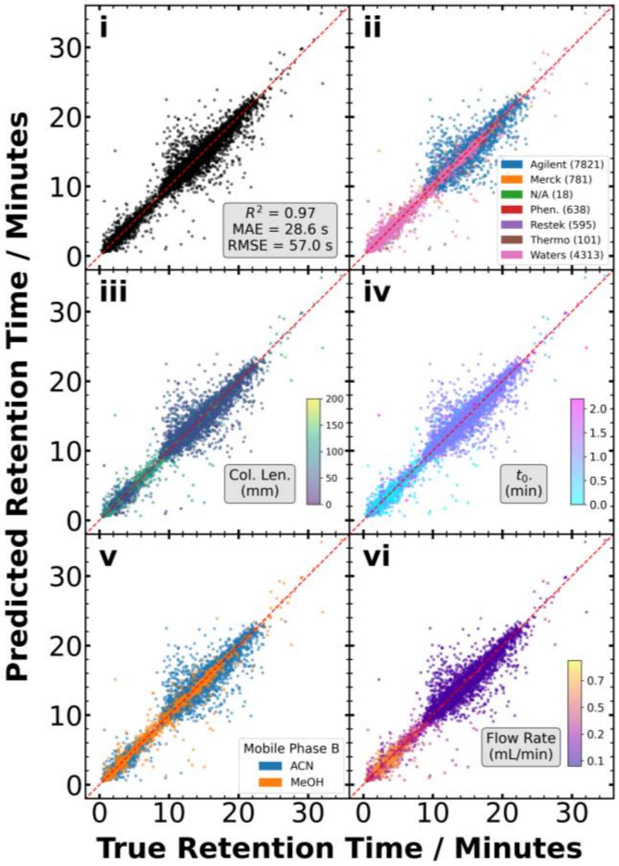

**Figure 3.** Test set analysis for the best-performing Graphormer-RT models for RP predictions showing performance **(i)** as a function of chromatographic conditions including manufacturer **(ii)**, column length **(iii)**, $t_0$ **(iv)**, organic phase composition **(v)**, and flow rate **(vi)**

confidence that Graphormer-RT is generalizing from chromatographic first principles. In contrast, Graphormer-RT fails to generalize (MAE = 596.8 ± 12.7 s) for method 0029.[35] This poor performance likely arises owing to the complexity of the gradient profile for method 0029 (see Figure S3), and it suggests that our formalism for gradient profiles may need to be improved for specific cases.

**Table 2.** Summary of model performance on held-out external validation sets. The number of cross validation folds (c) and the number of RTs per method (n) are given.

| **Reversed Phase Method** | **0029** [35]<br>**(*n* = 47)** | **0127** [34]<br>**(*n* = 93)** | **0275** [33]<br>**(*n* = 75)** |
|---|---|---|---|
| **External Test MAE (s)** | 596.8 ± 12.7 | 43.9 ± 3.3 | 42.7 ± 6.2 |
| **Column Type** | RP | RP | RP |
| **Mobile phase B** | ACN | MeOH | MeOH |
| **# Inflection Points** | 3 | 2 | 2 |
| **Manufacturer** | Waters | Merck | Phenomenex |
| **HILIC Method** | **0103** [34]<br>**(*n* = 70)** | **0283** [33]<br>**(*n* = 74)** | **0375** [16]<br>**(*n* = 59)** |
| **External Test MAE (s)** | 83.7 ± 7.3 | 66.6 ± 7.0 | 25.5 ± 2.5 |
| **Column Type** | HILIC | HILIC | HILIC |
| **Mobile phase A** | ACN | MeOH | ACN |
| **# Inflection Points** | 2 | 2 | 2 |
| **Manufacturer** | HILICON | Phenomenex | Thermo. |

To further explore the value of method-specific training, we proportionally reincorporated the held-out external methods back into the training-validation-test splits. Although an imperfect comparison to the "true" external evaluation due to differences in test set size (*n = 215 vs. n = 22*), this experiment provided a coarse estimation of the improvement offered by training on specific methods. For LC method 0029, introducing examples to the training set markedly improved test performance (by nearly an order of magnitude; MAE$_{incorporated}$ = 65.6 ± 21.1 s versus MAE$_{held\text{-}out}$ = 596.8 ± 12.7 s). Although the results for method 0029 are still worse than the average method performance, this outcome indicates that Graphormer-RT has learned the scaling factors for method 0029 reasonably well. From the experimental perspective, researchers that are employing complex or uncommon LC methods could consider "calibrating" Graphormer-RT by incorporating measured RTs from a calibration set into the training library.

**HILIC Results.** Our best performing HILIC model achieved MAE = 42.4 ± 2.9 s (*c = 5*, see **Table 1**). This model, which was first finetuned on the RP dataset, was trained using a library of 4,373 RTs associated with 49 HILIC methods. Interestingly, the improvement obtained from transfer learning was relatively small; uninitialized models achieve MAE = 44.5 ± 1.3 s. This result suggests that the RP and HILIC prediction tasks may be sufficiently dissimilar that transfer learning is not useful or that the HILIC training set size is sufficiently large and diverse to enable generalization. Figure S4 provides correlation diagrams for HILIC predictions, like those discussed earlier for RP methods (*i.e.*, Figures 3). In contrast to the model predictions for RP methods, MAEs for HILIC method predictions are not heteroscedastic. Whether this is a result of effective generalization for HILIC methods by Graphormer-RT or due to the implicit properties of the dataset is not known; a larger HILIC RT library is needed to investigate this further. That said, Graphormer-RT's performance on held-out external HILIC test sets does inspire some confidence in model generalization. As shown in **Table 2**, Graphormer-RT achieves MAEs of 83.7 ± 7.3 s, 66.6 ± 7.0 s, and 25.5 ± 2.5 s for RepoRT HILIC methods 0103,[34] 0283,[33] and 0375,[16] respectively. Regarding method 0375, Graphormer-RT predictions are more accurate by almost 20 s than the average global performance for methods that were included in the training set (MAE = 25.5 ± 2.5 s, see **Table 2**). A possible explanation for this relatively good performance could be that data for a very similar HILIC method is present in the training set. Given a sufficiently diverse set of HILIC methods, we expect that Graphormer-RT is expressive enough to predict RTs for novel methods. To the best of our knowledge, this is the first model that can accurately predict RTs for small molecule HILIC methods, as well as the first framework to generalize HILIC predictions across multiple methods.

**Feature Engineering.** Previous work has shown that removing descriptors can have a positive impact on model performance for learned node embeddings.[28,36] In some cases, certain descriptors may prove redundant or counterproductive to learning.[28,36,37] For all features used in this study, we perform ablations to identify such behavior. Tables S1-3 show all the encodings considered in this study and the final "pruned" set of descriptors that maximize model performance. For descriptors whose removal improved model performance, we propose possible explanations for why they are detrimental to model learning (see Supplementary Tables S1-3). For example, ablating the pH of the organic phase improved model performance. This may be because changes in organic phase pH may not have sufficient time to induce changes in chromatographic interactions; most compounds will have eluted before the organic phase becomes dominant (>50%, see Figure 1). For HILIC models, we found that including the concentration of the additives that are present in the mobile phase instead of simply signifying their presence with one hot encoding had a large, positive effect on performance – but this same behavior was not observed for RP predictions. This improvement is likely because the relative concentration of additives (*e.g.*, acidic or basic buffers) will influence the extent of hydrogen bonding occurring between analytes and the HILIC chromatography stationary phases.[38] For node (*i.e.*, atom) features, we found that removal of atomic mass and formal charge improved model performance. Atomic mass is likely redundant for RT prediction, since this information can be inferred from atomic number and the associated scaling factor is likely less important in this context than it might be in others (*e.g.*, IR frequency predictions).[28] Furthermore, the inclusion of partial charge, which encodes effects that are not well-described by the formal charge (*e.g.*, resonance effects), likely makes the formal charge descriptor redundant.

**Ablation Studies.** To test and justify architectural design, we performed ablation studies on the best performing RP prediction model using a single test split. This approach, which mitigates the runtime limitations associated with performing many ablation experiments, aligns with the protocols employed in previous works.[26,28] Even using a single test split, ablation studies provide insights regarding the important of specific features on the model performance.

Our best performing model (see **Table 3**, Ablation 1) employed a global chromatography node with learned embeddings that were generated from gradient features and column features. These features were passed into two separate int/float encoders, achieving a MAE = 28.6 s. The addition of the calculated Tanaka/HSMB parameters decreased model accuracy slightly to MAE = 29.3 s (see **Table 3**, Ablation 2),[39,40] suggesting that these parameters are made redundant by Graphormer-RT's derived understanding of chromatographic separation. The inclusion of a singular chromatography encoder instead of two divided by domain (integer and float) also slightly decreased model accuracy, yielding MAE = 29.9 s (see **Table 3**, Ablation 3). It may be that mixing of encoding types obfuscates learning. We also tested ablations that removed gradient features (*e.g.*, $\%B_{start}$, inflection points, solvent/pH/additives of both mobile phases) and column features (*e.g.*, diameter, particle size, length, $t_0$, temperature, manufacturer, and United States Pharmacopeial (USP) code). Removal of gradient features (see **Table 3**, Ablation 5) resulted in a more substantial reduction in accuracy, MAE = 33.4 s, than did removal of column features (MAE = 32.0 s; see **Table 3**, Ablation 4). These results suggest that gradient information may be slightly more important than column information, but both sets of features are needed for optimal model performance. We also tested the removal of all empirical descriptions of the chromatography, instead labeling each method as a discrete class. These encodings were passed through the chromatography encoder and achieved a MAE = 50.9 s (see **Table 3**, Ablation 6). This large decrease in accuracy compared to that of the best model demonstrates the value of including descriptive empirical method descriptions. If two methods have similar gradients with different mobile phases, models can use the similarity in those empirical descriptions to learn by analogy and more effectively generalize from chromatographic first principles. Finally, we investigated an ablation whereby we removed all method information (see **Table 3**, Ablation 7). Even without chromatography descriptors, Graphormer-RT yielded a MAE = 76.8 s. Removal of the chromatographic descriptors is tantamount to limiting learning to only task *(i)* (*i.e.*, relative molecular interactions), and the fact that even coarse predictions are possible demonstrates the ability of Graphormer-RT to learn interactions from molecular graph structure.

**Table 3.** Ablation study (*c* = 1) results for the best performing test split of the Graphormer-RT RP model. All training-validation-test splits are uniform across ablations. A checkmark indicates that the feature is present in the trained model.

| # | Chrom. Node | Grad. Feat. | Col. Feat. | Int/Float Encoders | Tanaka /HSMB | Test MAE (s) |
|---|---|---|---|---|---|---|
| **1.** | ✓ | ✓ | ✓ | ✓ |  | **28.6** |
| **2.** | ✓ | ✓ | ✓ | ✓ | ✓ | 29.3 |
| **3.** | ✓ | ✓ | ✓ |  |  | 29.9 |
| **4.** | ✓ | ✓ |  | ✓ |  | 33.4 |
| **5.** | ✓ |  | ✓ | ✓ |  | 32.0 |
| **6.** | ✓ |  |  |  |  | 50.9 |
| **7.** |  |  |  |  |  | 76.8 |

## Conclusions.

This study demonstrates the application of graph transformers in predictive method-independent models of RP and HILIC chromatography. After filtering the RepoRT dataset to improve method standardization,[16] we develop formalisms to describe the gradient and column metadata for machine learned graph embeddings. Along with the chromatographic method information, we employed SMILES strings as input. The final RP dataset contained 142,688 RTs associated with 191 distinct methods and the final HILIC dataset contained 4,373 RTs associated with 49 methods. A key addition was a set of "pre-graph" encoding neural networks designed to create a learned representation of gradient and column information. This information was fed into a flexible global node that was connected to all other nodes (*viz.* atoms) in the graph via a special edge type. This feature ensures that Graphormer-RT considers the column and gradient information *in the context* of molecular structure. Our best performing RP prediction method achieved a MAE = 29.3 ± 0.6 s, a meaningful improvement over previous state-of-the-art models, which employ only a single chromatographic method for training (compared to 191 here).[17] Our best performing HILIC model achieved MAE = 42.4 ± 2.9 s. To the best of our knowledge, Graphormer-RT is the first "method-independent" prediction of RP and HILIC RTs.

Using the rich RepoRT chromatographic metadata, we explored the relative performance of Graphormer-RT as a function of chromatographic parameters. For RP methods, Graphormer-RT performs best for methods that employ MeOH as the organic phase rather than ACN, low flow rates, low $t_0$, and Waters columns. HILIC methods showed no obvious bias in model performance with respect to chromatographic conditions, possibly due to the relatively small dataset size. Ablation studies showed that the best performing Graphormer-RT architecture consisted of a split (by datatype) “chromatographic” encoder. We also found that gradient features were more important for producing accurate model predictions than column features, though both are required for optimal model performance. Finally, although removing empirical chromatographic method data diminished model performance, Graphormer-RT still performed reasonably well, possibly due to its robust understanding of how molecular structure affects interaction strength.

To explore model generalizability, we held-out six methods (*i.e.*, 3 RP, 3 HILIC) from our training dataset. For four of these six methods, we find that Graphormer-RT generalizes reasonably well, yielding MAEs of less than one minute. However, Graphormer-RT completely failed to generalize for a complex gradient method (MAE = 596.8 ± 12.7 s), suggesting that a more complex encoding scheme may be needed in such instances. Encouragingly, though, including 80% of the RTs for this method in the training set led to a substantial improvement in model accuracy (MAE = 65.6 ± 21.0 s). This result suggests that researchers may be able to “calibrate” Graphormer-RT for more complex or uncommon chromatographic methods if examples are incorporated into training libraries.

To utilize the predicted RT predictions outside of a single laboratory/method framework for applications such as automated LC-MS$^{(2)}$ metabolite annotation, a “method-independent” model for RT prediction is needed. We achieved this goal using a novel graph transformer with contextual encodings for chromatographic metadata that generalizes well across diverse methods, including those that the model has not seen in training. We also demonstrate how this framework can be extended to small molecule HILIC chromatography. Although this study presents a large step forward in the generalizable predictions of RTs, more work can be performed to further generalize and optimize model performance. While we have demonstrated that Graphormer-RT can generalize to a variety of chromatographic conditions, training data for RTs measured in other conditions (*e.g.*, mobile phase of isopropanol, acetone) are needed for the model to meet the experimental chromatographic landscape. There are specific methods for which Graphormer-RT fails to generalize (*e.g.*, complex gradients). Addressing these will require expanding the diversity of training data or expanding encoding formalisms. Experimental LC-MS$^{(2)}$ has proven to be extremely powerful as a tool for separating, identifying, and annotating “dark” metabolites and other novel samples. With our contribution to “method-independent” predictions of RTs, we take a meaningful step towards the automated *in silico* annotation of unknown structures.

## ASSOCIATED CONTENT

## DATA AND SOFTWARE AVAILABILTY

Data is freely available in the RepoRT Github.[16] We also include a list of the methods used in our dataset after filtering. Code for this project will be made available online alongside Graphormer-RT code at https://github.com/HopkinsLaboratory/Graphormer-RT

## AUTHOR INFORMATION

Corresponding Author

*W. Scott Hopkins, shopkins@uwaterloo.ca

### Author Contributions

CMKS built the model architecture, performed all experiments, and wrote the manuscript. EN advised on encoding schemes/formalisms and helped prepare the manuscript. WSH was responsible for the concept and funding, provided experimental guidance, and aided in manuscript preparation.

### Funding Sources

CMKS acknowledges financial support from NSERC in the form of a Canadian Graduate Scholarship (Doctoral). WSH acknowledges funding from the Canadian Foundation for Innovation (CFI), Ontario Research Fund (ORF), Mitacs, and Natural Sciences and Engineering Research Council (NSERC) of Canada in the form of Discovery and Alliance grants.

## REFERENCES

(1) Taylor, P.; Nielsen, P. A.; Trelle, M. B.; Horning, O. B.; Andersen, M. B.; Vorm, O.; Moran, M. F.; Kislinger, T. Automated 2D Peptide Separation on a 1D Nano-LC-MS System. *J Proteome Res* **2009**, *8* (3), 1610–1616. https://doi.org/10.1021/PR800986C/.

(2) Johnston, J. J.; Draper, W. M.; Stephens, R. D. LC—MS Compatible HPLC Separation for Xenobiotics and Their Phase I and Phase II Metabolites: Simultaneous Anion Exchange and Reversed-Phase Chromatography. *J Chromatogr Sci* **1991**, *29* (12), 511–516. https://doi.org/10.1093/CHROMSCI/29.12.511.

(3) Domingo-Almenara, X.; Guijas, C.; Billings, E.; Montenegro-Burke, J. R.; Uritboonthai, W.; Aisporna, A. E.; Chen, E.; Benton, H. P.; Siuzdak, G. The METLIN Small Molecule Dataset for Machine Learning-Based Retention Time Prediction. *Nature Communications 2019 10:1* **2019**, *10* (1), 1–9. https://doi.org/10.1038/s41467-019-13680-7.

(4) Kanu, A. B. Recent Developments in Sample Preparation Techniques Combined with High-Performance Liquid Chromatography: A Critical Review. *J Chromatogr A* **2021**, *1654*. https://doi.org/10.1016/J.CHROMA.2021.462444.

(5) Stanstrup, J.; Neumann, S.; Vrhovšek, U. PredRet: Prediction of Retention Time by Direct Mapping between Multiple Chromatographic Systems. *Anal Chem* **2015**, *87* (18), 9421–9428. https://doi.org/10.1021/ACS.ANALCHEM.5B02287/.

(6) Creek, D. J.; Jankevics, A.; Breitling, R.; Watson, D. G.; Barrett, M. P.; Burgess, K. E. V. Toward Global Metabolomics Analysis with Hydrophilic Interaction Liquid Chromatography-Mass Spectrometry: Improved Metabolite Identification by Retention Time Prediction. *Anal Chem* **2011**, *83* (22), 8703–8710. https://doi.org/10.1021/AC2021823/.

(7) Strittmatter, E. F.; Kangas, L. J.; Petritis, K.; Mottaz, H. M.; Anderson, G. A.; Shen, Y.; Jacobs, J. M.; Camp, D. G.; Smith, R. D. Application of Peptide LC Retention Time Information in a Discriminant Function for Peptide Identification by Tandem Mass Spectrometry. *J Proteome Res* **2004**, *3* (4), 760–769. https://doi.org/10.1021/PR049965Y/.

(8) Bouwmeester, R.; Martens, L.; Degroeve, S. Comprehensive and Empirical Evaluation of Machine Learning Algorithms for Small Molecule LC Retention Time Prediction. *Anal Chem* **2019**, *91* (5), 3694–3703. https://doi.org/10.1021/ACS.ANALCHEM.8B05820/.

(9) Bach, E.; Schymanski, E. L.; Rousu, J. Joint Structural Annotation of Small Molecules Using Liquid Chromatography Retention Order and Tandem Mass Spectrometry Data. *Nature Machine Intelligence 2022 4:12* **2022**, *4* (12), 1224–1237. https://doi.org/10.1038/s42256-022-00577-2.

(10) Witting, M.; Böcker, S. Current Status of Retention Time Prediction in Metabolite Identification. *J. Sep. Sci.* **2020**, *43* (9–10), 1746–1754. https://doi.org/10.1002/jssc.202000060.

(11) Bouwmeester, R.; Gabriels, R.; Hulstaert, N.; Martens, L.; Degroeve, S. DeepLC Can Predict Retention Times for Peptides That Carry As-yet Unseen Modifications. *Nature Methods 2021 18:11* **2021**, *18* (11), 1363–1369. https://doi.org/10.1038/s41592-021-01301-5.

(12) Liu, Y.; Yang, Y.; Chen, W.; Shen, F.; Xie, L.; Zhang, Y.; Zhai, Y.; He, F.; Zhu, Y.; Chang, C. DeepRTAlign: Toward Accurate Retention Time Alignment for Large Cohort Mass Spectrometry Data Analysis. *Nature Communications 2023 14:1* **2023**, *14* (1), 1–12. https://doi.org/10.1038/s41467-023-43909-5.

(13) Giese, S. H.; Sinn, L. R.; Wegner, F.; Rappsilber, J. Retention Time Prediction Using Neural Networks Increases Identifications in Crosslinking Mass Spectrometry. *Nature Communications 2021 12:1* **2021**, *12* (1), 1–11. https://doi.org/10.1038/s41467-021-23441-0.

(14) Haddad, P. R.; Taraji, M.; Szücs, R. Prediction of Analyte Retention Time in Liquid Chromatography. *Anal Chem* **2021**, *93* (1), 228–256. https://doi.org/10.1021/ACS.ANALCHEM.0C04190/.

(15) Osipenko, S.; Botashev, K.; Nikolaev, E.; Kostyukevich, Y. Transfer Learning for Small Molecule Retention Predictions. *J Chromatogr A* **2021**, *1644*, 462119. https://doi.org/10.1016/J.CHROMA.2021.462119.

(16) Kretschmer, F.; Harrieder, E. M.; Hoffmann, M. A.; Böcker, S.; Witting, M. RepoRT: A Comprehensive Repository for Small Molecule Retention Times. *Nature Methods 2024 21:2* **2024**, *21* (2), 153–155. https://doi.org/10.1038/s41592-023-02143-z.

(17) Osipenko, S.; Nikolaev, E.; Kostyukevich, Y. Retention Time Prediction with Message-Passing Neural Networks. *Separations* **2022**, *9* (10), 291. https://doi.org/10.3390/SEPARATIONS9100291.

(18) Vik, D.; Pii, D.; Mudaliar, C.; Nørregaard-Madsen, M.; Kontijevskis, A. Performance and Robustness of Small Molecule Retention Time Prediction with Molecular Graph Neural Networks in Industrial Drug Discovery Campaigns. *Scientific Reports 2024 14:1* **2024**, *14* (1), 1–8. https://doi.org/10.1038/s41598-024-59620-4.

(19) Yang, Q.; Ji, H.; Lu, H.; Zhang, Z. Prediction of Liquid Chromatographic Retention Time with Graph Neural Networks to Assist in Small Molecule Identification. *Anal Chem* **2021**, *93* (4), 2200–2206. https://doi.org/10.1021/ACS.ANALCHEM.0C04071.

(20) Xu, H.; Lin, J.; Zhang, D.; Mo, F. Retention Time Prediction for Chromatographic Enantioseparation by Quantile Geometry-Enhanced Graph Neural Network. *Nature Communications 2023 14:1* **2023**, *14* (1), 1–15. https://doi.org/10.1038/s41467-023-38853-3.

(21) Xu, H.; Lin, J.; Zhang, D.; Mo, F. Retention Time Prediction for Chromatographic Enantioseparation by Quantile Geometry-Enhanced Graph Neural Network. *Nature Communications 2023 14:1* **2023**, *14* (1), 1–15. https://doi.org/10.1038/s41467-023-38853-3.

(22) Drvodelic, M.; Gong, M.; Webb, A. I. GraphRT: A Graph-Based Deep Learning Model for Predicting the Retention Time of Peptides. **2024**.

(23) Fedorova, E. S.; Matyushin, D. D.; Plyushchenko, I. V.; Stavrianidi, A. N.; Buryak, A. K. Deep Learning for Retention Time Prediction in Reversed-Phase Liquid Chromatography. *J Chromatogr A* **2022**, *1664*, 462792. https://doi.org/10.1016/J.CHROMA.2021.462792.

(24) Zaretckii, M.; Bashkirova, I.; Osipenko, S.; Kostyukevich, Y.; Nikolaev, E.; Popov, P. 3D Chemical Structures Allow Robust Deep Learning Models for Retention Time Prediction. *Digital Discovery* **2022**, *1* (5), 711–718. https://doi.org/10.1039/D2DD00021K.

(25) Osipenko, S.; Bashkirova, I.; Sosnin, S.; Kovaleva, O.; Fedorov, M.; Nikolaev, E.; Kostyukevich, Y. Machine Learning to Predict Retention Time of Small Molecules in Nano-HPLC. *Anal Bioanal Chem* **2020**, *412* (28), 7767–7776. https://doi.org/10.1007/S00216-020-02905-0/.

(26) Ying, C.; Cai, T.; Luo, S.; Zheng, S.; Ke, G.; He, D.; Shen, Y.; Liu, T.-Y. Do Transformers Really Perform Bad for Graph Representation? *NIPS'21: Proceedings of the 35th International Conference on Neural Information Processing Systems* **2021**, 28877–28888. https://doi.org/https://arxiv.org/abs/2106.05234.

(27) Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, Ł.; Polosukhin, I. Attention Is All You Need. *Adv Neural Inf Process Syst* **2017**, *2017-December*, 5999–6009.

(28) Stienstra, C. M. K.; Hebert, L.; Thomas, P.; Haack, A.; Guo, J.; Hopkins, W. S. Graphormer-IR: Graph Transformers Predict Experimental IR Spectra Using Highly Specialized Attention. *J Chem Inf Model* **2024**, *64* (12), 4613–4629. https://doi.org/10.1021/ACS.JCIM.4C00378.

(29) Das, A.; Shuaibi, M.; Palizhati, A.; Goyal, S.; 1→3, A. G.; Kolluru, A.; Lan, J.; Rizvi, A.; Sriram, A.; Wood, B.; Parikh, D.; Ulissi, Z.; Zitnick, C. L.; Ke, G.; Zheng, S.; Shi, Y.; He, D.; Liu, T.-Y.; Ying, C.; You, J.; He, Y.; Grigoriev, R.; Lukin, R.; Yarullin, A.; Faleev, M.; Kiela, D.; Ciccone, M.; Caputo, B. The Open Catalyst Challenge 2021: Competition Report. *Proceedings of Machine Learning Research*. PMLR July 20, 2022, pp 29–40. https://proceedings.mlr.press/v176/das22a.html (accessed 2023-07-20).

(30) Hu, W.; Fey, M.; Ren, H.; Nakata, M.; Dong, Y.; Leskovec, J. OGB-LSC: A Large-Scale Challenge for Machine Learning on Graphs. **2021**.

(31) Irwin, J. J.; Tang, K. G.; Young, J.; Dandarchuluun, C.; Wong, B. R.; Khurelbaatar, M.; Moroz, Y. S.; Mayfield, J.; Sayle, R. A. ZINC20 - A Free Ultralarge-Scale Chemical Database for Ligand Discovery. *J Chem Inf Model* **2020**, *60* (12), 6065–6073. https://doi.org/10.1021/ACS.JCIM.0C00675/.

(32) Folberth, J.; Begemann, K.; Jöhren, O.; Schwaninger, M.; Othman, A. MS2 and LC Libraries for Untargeted Metabolomics: Enhancing Method Development and Identification Confidence. *Journal of Chromatography B* **2020**, *1145*, 122105. https://doi.org/10.1016/J.JCHROMB.2020.122105.

(33) Souihi, A.; Mohai, M. P.; Palm, E.; Malm, L.; Kruve, A. MultiConditionRT: Predicting Liquid Chromatography Retention Time for Emerging Contaminants for

a Wide Range of Eluent Compositions and Stationary Phases. *J Chromatogr A* **2022**, *1666*, 462867. https://doi.org/10.1016/J.CHROMA.2022.462867.

(34) Folberth, J.; Begemann, K.; Jöhren, O.; Schwaninger, M.; Othman, A. MS2 and LC Libraries for Untargeted Metabolomics: Enhancing Method Development and Identification Confidence. *Journal of Chromatography B* **2020**, *1145*, 122105. https://doi.org/10.1016/J.JCHROMB.2020.122105.

(35) Low, D. Y.; Micheau, P.; Koistinen, V. M.; Hanhineva, K.; Abrankó, L.; Rodriguez-Mateos, A.; da Silva, A. B.; van Poucke, C.; Almeida, C.; Andres-Lacueva, C.; Rai, D. K.; Capanoglu, E.; Tomás Barberán, F. A.; Mattivi, F.; Schmidt, G.; Gürdeniz, G.; Valentová, K.; Bresciani, L.; Petrásková, L.; Dragsted, L. O.; Philo, M.; Ulaszewska, M.; Mena, P.; González-Domínguez, R.; Garcia-Villalba, R.; Kamiloglu, S.; de Pascual-Teresa, S.; Durand, S.; Wiczkowski, W.; Bronze, M. R.; Stanstrup, J.; Manach, C. Data Sharing in PredRet for Accurate Prediction of Retention Time: Application to Plant Food Bioactive Compounds. *Food Chem* **2021**, *357*, 129757. https://doi.org/10.1016/J.FOODCHEM.2021.129757.

(36) Masters, D.; Dean, J.; Klaser, K.; Li, Z.; Maddrell-Mander, S.; Sanders, A.; Helal, H.; Beker, D.; Rampášek, L.; Beaini, D. GPS++: An Optimised Hybrid MPNN/Transformer for Molecular Property Prediction. **2022**.

(37) Calude, C. S.; Longo, G. The Deluge of Spurious Correlations in Big Data. *Found Sci* **2017**, *22* (3), 595–612. https://doi.org/10.1007/S10699-016-9489-4.

(38) Buszewski, B.; Noga, S. Hydrophilic Interaction Liquid Chromatography (HILIC)—a Powerful Separation Technique. *Anal Bioanal Chem* **2012**, *402* (1), 231. https://doi.org/10.1007/S00216-011-5308-5.

(39) Snyder, L. R.; Dolan, J. W.; Carr, P. W. The Hydrophobic-Subtraction Model of Reversed-Phase Column Selectivity. *J Chromatogr A* **2004**, *1060* (1–2), 77–116. https://doi.org/10.1016/J.CHROMA.2004.08.121.

(40) Kimata, K.; Iwaguchi, K.; Onishi, S.; Jinno, K.; Eksteen, R.; Hosoya, K.; Araki, M.; Tanaka, N. Chromatographic Characterization of Silica C18 Packing Materials. Correlation between a Preparation Method and Retention Behavior of Stationary Phase. *J Chromatogr Sci* **1989**, *27* (12), 721–728. https://doi.org/10.1093/CHROMSCI/27.12.721.

(41) Bonini, P.; Kind, T.; Tsugawa, H.; Barupal, D. K.; Fiehn, O. Retip: Retention Time Prediction for Compound Annotation in Untargeted Metabolomics. *Anal Chem* **2020**, *92* (11), 7515–7522. https://doi.org/10.1021/ACS.ANALCHEM.9B05765.

(42) Win, Z.-M.; M. Y. Cheong, A.; Scott Hopkins, W. Using Machine Learning To Predict Partition Coefficient (Log P) and Distribution Coefficient (Log D) with Molecular Descriptors and Liquid Chromatography Retention Time. *J Chem Inf Model* **2023**, *63* (7), 1906–1913. https://doi.org/10.1021/acs.jcim.2c01373.

(43) Shi, Y.; Zheng, S.; Ke, G.; Shen, Y.; You, J.; He, J.; Luo, S.; Liu, C.; He, D.; Liu, T.-Y. Benchmarking Graphormer on Large-Scale Molecular Modeling Datasets. **2022**.

(44) Li, X.; Zhang, J.; Wang, S.; Zhou, Q. Two-Stream Spatial Graphormer Networks for Skeleton-Based Action Recognition. *IEEE Access* **2022**, *10*, 100426–100437. https://doi.org/10.1109/ACCESS.2022.3206044.

(45) Hebert, L.; Chen, H. Y.; Cohen, R.; Golab, L. Qualitative Analysis of a Graph Transformer Approach to Addressing Hate Speech: Adapting to Dynamically Changing Content. **2023**.

(46) Mao, Z.; Zhang, R.; Xin, L.; Li, M.; Lei, X. Mitigating the Missing Fragmentation Problem in de Novo Peptide Sequencing with a Two Stage Graph-Based Deep Learning Model. **2023**. https://doi.org/10.21203/RS.3.RS-2593528/V1.

(47) RDKit: Open-Source Cheminformatics. *https://www.rdkit.org*.

(48) 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* **2019**, *59* (8), 3370–3388. https://doi.org/10.1021/ACS.JCIM.9B00237.

(49) Fred Agarap, A. M. Deep Learning Using Rectified Linear Units (ReLU). **2018**.

(50) White, A. D. Deep Learning for Molecules and Materials. *Living J Comput Mol Sci* **2021**, *3* (1), 1499–1499. https://doi.org/10.33011/LIVECOMS.3.1.1499.

(51) Paszke, A.; Gross, S.; Massa, F.; Lerer, A.; Bradbury, J.; Chanan, G.; Killeen, T.; Lin, Z.; Gimelshein, N.; Antiga, L.; Desmaison, A.; Köpf, A.; Yang, E.; DeVito, Z.; Raison, M.; Tejani, A.; Chilamkurthy, S.; Steiner, B.; Fang, L.; Bai, J.; Chintala, S. PyTorch: An Imperative Style, High-Performance Deep Learning Library. *Adv Neural Inf Process Syst* **2019**, *32*.

(52) Li, M.; Zhou, J.; Hu, J.; Fan, W.; Zhang, Y.; Gu, Y.; Karypis, G. DGL-LifeSci: An Open-Source Toolkit for Deep Learning on Graphs in Life Science. *ACS Omega* **2021**, *6* (41), 27233–27238. https://doi.org/10.1021/acsomega.1c04017.

(53) Kingma, D. P.; Ba, J. L. Adam: A Method for Stochastic Optimization. *3rd International Conference on Learning Representations, ICLR 2015 - Conference Track Proceedings* **2014**.

(54) Prechelt, L. Early Stopping - But When? **1998**, 55–69. https://doi.org/10.1007/3-540-49430-8_3.

## Methods

**Dataset.** In this study, we utilize the RepoRT dataset,[16] an evolving "machine learning ready" repository for retrieving RTs with chromatographic metadata. At the time of access, RepoRT contained 392 different methods, 172,416 total retention times for 94,788 unique molecules represented by SMILES strings. [16] RepoRT is a superset of a variety of commonly used retention time datasets including the METLIN SMRT,[3] Retip,[41] and PredRet datasets,[5] and it contains RTs measured on reversed phase (RP), HILIC, and amide columns. For each method included in the dataset, there is a rich description of the column metadata that includes time-variant gradient profiles, mobile phase additives, intrinsic column parameters, calculated Tanaka and HSMB column parameters,[16] among other descriptors of chromatographic conditions. While RepoRT dataset does perform preliminary filtering of retention times (error thresholds and void volumes), we employed additional filtering steps to maximize the learnability of the dataset and the representation of chromatographic parameters.

**Reversed Phase (RP) Methods.** Mobile phases in RepoRT include water ($H_2O$), MeOH, ACN, isopropanol (iPrOH), and acetone (ACE).[16] Since the methods using ACE and IPA only describe 293 and 378 unique retention times, respectively, we exclude them from our dataset. We also exclude all gradients containing more than two distinct mobile phases. We ensure that all mobile phases are described such that the solvent A is the aqueous component, and solvent B is the organic. All gradients with a non-constant flow rate are also excluded from our training dataset. All gradients with $t_0$ greater than 3 minutes were excluded from our dataset because of the high variability associated with the measured RTs. As has been performed in other studies using the METLIN SMRT dataset,[17,42] we exclude all molecules not retained on the column because of the large error associated with the dataset bimodality. Our final, pruned, RepoRT dataset consisted of 191 unique methods, totaling 142,668 retention times, and 89,643 unique SMILES strings. Summary statistics for the chemical and chromatographic makeup of this dataset are available in Figures S10-26.

**Gradient Descriptions.** The composition and gradient of a mobile phase in an LC method is highly relevant to the prediction of an arbitrary molecule's retention time. Our previous work has shown that for deep learning, it is important to balance providing models a sufficiently rich initial human-engineered description of the parameter space and not over-describing the problem based on human intuition in a way that gives the model an inflexible and overly specific initialization.[28] This is particularly true for mobile phase gradients which are described by complex, parameterized, piecewise time function that might contain extemporaneous information.

We attempt to distill the RepoRT descriptions of gradient profiles into the simplest description that contains all useful gradient information. Since all filtered methods contain only two solvents at a constant flow rate, we can express the gradient as % solvent B as a function of time (see Figure 1). To extract the changes in solvent composition and polarity as a function of time, we identify all inflection points in the curve between the minimum and maximum %B regions, and include these coordinates (time(min), %B) as input parameters (see Figure 1). We also include the starting solvent composition (%B) as an input parameter. We limit our dataset to only include gradients with at most three of these inflection points, which represents most of the RepoRT dataset. No gradient information is included after the system reaches $\%B_{max}$, since this is typically the column re-equilibration phase, where functionally all compounds have already eluted.  As shown in Figure S27, certain gradients reach a %B near 100% and then immediately (typically after ~0.1 minutes) jump to %B = 100. We found that the inclusion of two inflection points in these instances diminished model accuracy, so we remove the first of the two points. We also include one-hot encodings describing the presence or absence of mobile phase additives such as formic acid, triethylamine (TEA), mandelic acid, or ammonium acetate (among others) in both mobile phases. We provide the pH of both mobile phases as a float encoding. For all float features, we attempted to normalize values to the dataset distribution but found that it had either a null or detrimental effect to performance. A complete summary of all mobile phase and gradient descriptors, data types, and domains are present in the supplementary information Tables S1 and S2.

**Intrinsic Column Descriptors.** It is well known that column dimensions, manufacturer, and type influence the retention times of all compounds that transit that column. In the RepoRT dataset, these parameters were collected and standardized from publications and manufacturer libraries, meaning that this data could be used as descriptors for model learning. For the columns themselves, we include descriptions of several important column parameters including the column dimensions (*e.g.*, length inner diameter, particle size) in standardized units, the dead/void volume in minutes, and the temperature in degrees Kelvin. We provide one hot description of the column manufacturer (*e.g.*, Thermofisher, Phenomenex), and USP code.[16] If a column parameter was not provided in the original study, we encode that value as a zero.

**HILIC Methods.** Identical encoding schemes and filtering methods were utilized for HILIC methods as were for the RP methods, ensuring that methods had a maximum of three inflection points, binary solvent composition, and constant flow rate. The primary difference between HILIC methods and RP methods in our dataset is that gradients are typically described as starting at $\%B \approx 100$ % (i.e., the organic phase), with increasing contributions from the aqueous phase. This profile produces gradients that are effectively "upside-down" compared to those encoded in our RP models (see Figure 1). To maximize similarity to the gradient encodings in our RP models for the purposes of transfer learning, we invert the labels (*e.g.*, A, B) for the mobile phase gradients such that the gradient starts at a %B minimum and matches the shape of the RP profiles. While this has the consequence of inverting the character of mobile phases A and B (*viz.* the organic and aqueous character), we believe that this difference is more easily learned as compared to a complete shift of all the gradient profile encodings. Additionally, in most of the HILIC methods, the organic phases (typically ACN) are not pure and contain 5-10% $H_2O$. Since these concentrations are relatively consistent, we approximate the organic phase as 100% the organic solvent, for ease of encoding and because the small volume of water should be learned implicitly by the model. We found that including the concentration of the additives present in the mobile phase instead of just a one hot encoding had a large, positive effect in performance not observed for RP predictions. After filtering and consolidation, our HILIC library consists of 49 methods, totaling 4,373 retention times.

**External Datasets.** A consequence of dividing the entire library of retention times into the training, validation, and testing splits is that it is very likely that in training, the model will "see" the embeddings of a method before the generalizability is tested. In other words, most methods will appear in the training sets before evaluation in the test set. From the machine learning perspective this causes no data leakage because models will never see the ground truth retention times. However, if we choose to understand learning RTs as a combination of the two tasks discussed in the *Introduction* section, this effect will likely result in some memorization/leakage with respect to task (*ii)*, the chromatographic method rescaling. In this work, test the "true" generalizability of Graphormer-RT by using methods *and* retention times that the models have never seen. For experimental pre-screening, this measures how effectively Graphormer-RT can predict retention times for completely "new" methods and require no experimental measurements to accompany retention time prediction. We remove six (3 RP, 3 HILIC) "representative" methods from our dataset from all training workflows. Upon testing, we report the error of these methods separately to the testing dataset (which contains methods that could have been seen in training/validation) to provide a more complete measure of model generalizability. For RP systems, we choose methods 0029 (*n=47),*[35] 0127 (*n=93),*[34] and 0275 (*n=75)*[33] as denoted in RepoRT. For HILIC datasets we use methods 0103 (*n=70),*[34] 0283 (*n=74),*[33] and 0375 (*n=59)*.[16] These methods provide a diverse but representative set of columns, manufacturers, and gradients to maximize the diversity of our testing protocols. Complete descriptions of these chromatographic methods are available in Figures S1-S6.

**Model Architecture: Transformers.** We choose Graphormer as our architecture because of its success in a variety of tasks and have been shown to outpace MPNNs, which have previously been the state-of-the-art model for many (bio)cheminformatics tasks.[26,43–46] Transformers have been very successful because of their ability to describe relational data with highly contextual learned aggregations of the input representations. Attention mechanisms are largely responsible for this expressiveness, where models can assign learned degrees of relative importance between input tokens (atoms, for our purposes) in the input sequence. These higher order relationships between tokens are aggregated to generate dense representations with highly contextual of long-ranged interactions.[27] While used with great success in natural language processing (NLP), transformers were long thought to perform poorly when describing non-linear, graph structured data.

The advent of Graphormer provided a major step in showing that transformers could be extended to this graph structured data. A detailed description of the mathematical underpinnings of Graphormer's novelties can be found in references [26] and [28]. Graph structured data can include social networks, recommendation systems, or molecules. Expressive descriptions of these systems require a detailed consideration of the structural representation of entities that can be described by nodes and edges. In cheminformatics, this description is akin to expressing the fundamental rules of bonding and chemical stability embedding in molecular structure. Graphormer describes these phenomena by utilizing a multiheaded self-attention mechanism which allows for contextual descriptions that represent contextual relationships between nodes (atoms) and edges (bonds). Graphormer also uses a robust consideration of the input graph structure to ensure a rich consideration of the complex relationships described in graphs. Edge and spatial encodings (that are indexed to the shortest distance between connected node pairs) are passed to the attention mechanism, allowing Graphormer to learn contextual relationships that depend on graph structure and include scaling based on graph distance. Graphormer also utilizes a *global receptive field*, meaning that all nodes in a graph consider relationships with all other nodes in the graph in the attention mechanism.[26] As such, the embeddings generated by Graphormer might consider important long range chemical interactions in a molecule (*e.g.*, hydrogen bonding).

**Graph Feature Encoding.** In using Graphormer to predict infrared spectra, we explored the best way to encode atomic and edge descriptors into the graph structure. A detailed summary of these methods is available in reference [28]. The emphasis of this work was on understanding how to balance a knowledgeable model initialization (by using descriptors like hybridization, Gasetgier partial charge, *etc.*)[47] and providing a flexible description of the local chemical environment to not shoehorn model into an overly specific human-engineered description. We found that the optimal architecture involved using a learned feature encoder (**3** in Figure 2), where node features are projected into a fixed-size latent space using a small neural network. All hydrogens are described implicitly (as node features), with the exception for those that define stereocenters. We describe edge features using the combinatoric mapping scheme utilized in the Graphormer-IR study.[28] All node and edge features used to initialize these encoders can be found in Tables S1 and S2 with associated domains.

**Global Chromatography Node.** In our previous work,[28] we used a global graph node connected to all other nodes in the molecular graph with special edge types to encode global properties such as spectral phase (*i.e.*, gas phase, nujol mull, $CCl_4$, *etc).* These global features were expressed as one-hot encodings (analogous to atomic number) and were found to generate embeddings that reflected the global state *in the context of* the molecular structure itself. For IR spectral predictions, this structural context was useful for predictions of *bathochromic shifts*, where the solvent phase can cause varying shifts in specific IR frequencies.[28]

This contextualization of predictions inspires confidence that the use of a global graph node improves model understanding of emergent chemistry. For retention time prediction across different chromatographic setups, we believe that molecular context is of the upmost importance. Not all functional groups will interact identically with the stationary or mobile phase, and these interactions and they impact retention times may change, or scale differently based on chromatographic conditions. Mobile phase pH provides an illustrative example in this regard. For molecules with ionizable functional groups, a difference of 2 pH units in the mobile phase may represent (de)protonation of those functional groups, creating a dramatic difference in retention time driven primarily by ionic or hydrophobic interactions with a C18 column. For molecules without ionizable functional groups, the shift in pH will likely not be as meaningful. By describing these parameters by a *global chromatography node*, we allow Graphormer's attention mechanisms to derive relationships between the nodes in the graph (including the global node) that allows models to derive relationships between the molecular structure and these global features. If we simply use the traditional approach[48] and append these features to the latent dense "fingerprint" representation (see Figure 2), the decoding MLP loses ability to consider these properties in the *context* of the molecular structure because the tensor containing the chromatographic information loses its description of graph structure.

For spectral phase, we expressed the encodings as distinct classes analogous to a "new" atom type label that the model could appropriate to describe the changes in behavior. By mixing this atom type with a null description for all other node descriptors, we provide a sufficiently descriptive encoding for the global phase that allows the model to learn the different spectral encodings. However, for the prediction of RTs, there are many useful empirical descriptions that would be lost by consolidating methods into a single class. Additionally, these empirical chromatographic parameters describe fundamentally different chemical phenomena and as compared to those for the descriptions of atom (node) environments. As such, mixing these parameter spaces in a single node feature encoder will almost certainly impinge on model learning.

Given this understanding, we introduced two additional pre-graph encoders to perform learned projections of the column parameters into a distinct embedding space. Global nodes are labelled and their parameters (which are padded to have identical shapes as the node features) are passed to global chromatography feature encoders. We found that separating the integer (*viz.* one-hot) and float features (**1, 2** in Figure 2) into separate encoders improved model performance. These encoders are very similar in shape to the node feature encoder,[28] where column features are projected into a fixed-size latent space using two linear layers that project the number of features into an intermediate size of 256, then the embedding dimension (512 for the optimized model), using Rectified Linear Units (ReLUs) activation and dropout layers ($p = 0.05$). [49] Because this encoder has independent tunable parameters to the node feature encoder, the column parameters can be projected independently of the node features. All the benefits of the node feature encoder apply to the chromatography encoder as well, where a learned representation of column parameters maximizes the flexibility and informativity of the description passed to Graphormer without unnecessarily shoehorning model understanding by over-describing the input.

**Transfer Learning.** Transfer learning is a powerful tool inspired by the psychology of human learning that is widely utilized in NLP and machine learning. A detailed description of the applications and theory of transfer learning is available in reference [50]. In the same way that humans improve learning by using analogy, so too can deep learning frameworks improve performance by first being trained on a secondary task (perhaps with a larger dataset) that utilizes analogous mechanics to the primary task. While tasks may not be identical in scope, analogous skills can be applied in different contexts to accelerate learning.

In this study, we choose to treat the tasks of RP and HILIC chromatography as analogous but fundamentally different tasks. While they share an underlying analytical technique brand, the fundamental interactions that dictate the retention times of molecules are intrinsically different from one another and attempting to predict both at the same time with a single model is akin to pulling the models latent description of molecular graph in two opposite directions at once. However, due to the similarity of the principles and input descriptions of these chromatographic methods they present an excellent use case for the principles of transfer learning. As such, we attempt to transfer the knowledge learned from our RP model to improve learning of HILIC columns and see if it is sufficient overcome the relative dataset paucity.

**Model Training.** All models trained in this work are built on the original Graphormer architecture,[26] built on the PyTorch library.[51] We use our and graph feature encoders ( **3** in Figure 2) and chromatographic feature encoders (**1,2** in Figure 2) to create molecular graphs with node- (*i.e.*, atom) and edge- (*i.e.*, bond) features using the DGL-LifeSci package[52] with dense embeddings (**4** in Figure 2). These graphs are passed to Graphormer (**5** in Figure 2) to construct a set of learned node representations and the aggregate mCLS token. This molecular fingerprint is interpreted by the MLP (**3** in Figure 6), consisting of intermediate linear layers and activation functions, which predicts the final retention time. The best performing model contained eight Graphormer encoder layers with an embedding dimension of 512. A self-attention mechanism with 64 attention heads was employed with an attention dropout of 0.15. Adam was used as the optimizer with $\beta_1 = 0.9$, $\beta_2 = 0.999$, and $\epsilon = 10^{-8}$.[53] Reversed phase models were trained using an initial learning rate of $1.0 \times 10^{-4}$ and HILIC models used a learning rate of $1.5 \times 10^{-4}$. The MLP (decoder) consisted of four linear layers with an output of size 512, interleaved with ReLUs[49] as the activation function. The loss function used in training was the root mean squared error (RMSE). In total, the best performing model had 54,700,619 tunable parameters. All hyperparameters for the best performing model were determined by manual tuning using the validation loss. All models were trained for 250 epochs, with a batch size of 64 and early stopping.[54] All models utilized an 80–10–10 (%) training-validation-test split. All models were trained on a single NVIDIA GeForce RTX 4090 graphics card with 24 GB of VRAM. Training a model with RP retention times takes roughly 7.02 hours to complete and finetuning on HILIC retention times takes roughly 14.7 minutes. One of the major benefits of the machine learning approach for RT prediction is that once a model is trained, calculation of retention times is extremely rapid. Evaluation of the total RP retention time test set (n = 14,482) takes 3.4 minutes, which corresponds to roughly 14 ms per retention time prediction.