Beyond Group Additivity: Transfer Learning for Molecular Thermochemistry Prediction

Yannick Ureel[1], Florence H. Vermeire[2], Maarten K. Sabbe[1], Kevin M. Van Geem[1,*]

[1]Laboratory for Chemical Technology, Department of Materials, Textiles and Chemical Engineering, Ghent University, Technologiepark 125, 9052 Gent, Belgium
[2]Department of Chemical Engineering, KU Leuven, Celestijnenlaan 200F, 3001 Leuven, Belgium

[*] Corresponding author: Kevin.VanGeem@UGent.be, Technologiepark 125, 9052 Gent, Belgium;

Keywords: Thermodynamics, Hydrocarbons, Radical, Carbenium ion, SHapley Additive exPlanations

Abstract:

The accuracy of thermochemical prediction methods is strongly dependent on the size of the set of training data. Group additivity is an interpretable modeling strategy that can be developed from a limited dataset, but fails to consider delocalized molecular effects such as inductive stabilization, delocalized resonance stabilization, and steric effects. In contrast, machine learning allows the incorporation of these effects but requires an extensive amount of high-quality data. Therefore, a new transfer learning approach is proposed, uniting group additivity with machine learning. First, a machine learning model is pretrained on a large set of group additive predictions, after which it is refined on a limited high-quality dataset with transfer learning. The proposed approach was tested to predict the standard enthalpy of formation, standard molar entropy, and heat capacity of a wide range of hydrocarbons, hydrocarbon radicals, and carbenium ions. By using transfer learning, chemically accurate predictions for hydrocarbons, radicals, and carbenium ions could be obtained, drastically reducing the group additive error using less than 450 molecular datapoints per model. A SHapley Additive exPlanations analysis reveals that a data-efficient but interpretable transfer learning methodology is obtained, achieving chemically accurate predictions for a wide range of hydrocarbons.

1. Introduction

The accurate and fast prediction of molecular properties is an essential asset of chemical engineering, both for the development of new materials and molecules or the optimization of chemical processes [1, 2]. The thermochemistry of molecules in the gas phase, being the standard enthalpy of formation ($\Delta H^0_{f,298}$), the standard molar entropy (S^0_{298}), and heat capacity (C^0_p), are of special interest. These properties are essential to maintain thermodynamic consistency during first-principle kinetic model generation and the evaluation of energy balances for reactor and catalyst design [3, 4]. One modeling strategy for the prediction of these properties are group additive models.

Group additivity for thermochemical property prediction was originally developed by Benson et al. [5]. In this approach, a molecule is divided into several atom-centered groups and the predicted property is the sum of the thermochemical contributions of every group. The value of the contribution of every group, also called group additive value (GAV), is determined through regression of either experimental or ab initio values [5-7]. However, these GAVs can only account for interactions contained within one group, which can limit the accuracy of this approach. In the Benson approach, a group is a molecular substructure comprising of one atom with its corresponding neighboring atoms. On top of these groups, non-nearest neighbor interactions (NNI) can be defined which allow to incorporate localized effects beyond the range of a group. These GAVs and NNIs are summed up to obtain the group additive predictions.

Group additive models for molecular thermochemical prediction have been constructed for a wide range of gas phase molecules. An established experimental set of GAVs and NNIs for the standard enthalpy of formation for hydrocarbons, and oxygenates based on experimental values is the compilation of Cohen [7]. Sabbe et al. extended the group additivity approach for the

prediction of the standard enthalpy of formation of radicals and also for the prediction of the standard molar entropy and heat capacity of a wide range of hydrocarbons and radicals [6, 8]. GAVs have also been determined for property prediction of organosulfur compounds [9], oxygenated hydrocarbons [10], and aromatic hydrocarbons [11, 12]. Besides gas phase molecular properties, properties of ionic liquids or other organic molecules have also been thoroughly investigated with group additivity [13-17]. Acree et al. and Naef et al. determined GAVs of various physical molecular properties of ionic liquids and organic molecules such as the enthalpy of formation [18], heat of vaporization [14, 15], and vapor pressure [19]. The use of group additivity in chemical engineering is not limited to molecular property prediction. Group additive models have also been constructed for the prediction of activation energies and pre-exponential factors of various chemical reactions [20-25]. Furthermore, GAVs also exist for other properties such as viscosity [26, 27], surface tension [13, 27], threshold sooting index [28], solubility parameters [16, 29], and boiling and melting temperatures [30].

Group additive models offer many benefits, as they can be constructed with a limited amount of data, are interpretable as the contribution of every group to the property is determined, and allow a fast and reasonably accurate prediction. However, there are also limitations when it comes to the accuracy of group additive approaches. Non-localized effects such as steric effects, resonance and inductive stabilization cannot be incorporated into GAVs or NNIs. These effects can significantly influence the thermochemistry of molecules and accounting for them can thus further improve the prediction accuracy. Machine learning models can overcome these limitations as they allow to consider the complete molecule instead of well-localized groups [31, 32]. Chung et al. compared the performance of group additivity with machine learning for the prediction of solvation energies and found machine learning models superior in accuracy with datasets all larger than 6000 molecules [29]. However, these models lack interpretability,

and moreover require a lot of high-quality data, the latter being a major bottleneck for the applicability of machine learning for molecular property prediction. Hence, strategies have been developed to make machine learning less data intensive such as transfer learning [33-37] and physics-informed learning, or to aid data generation in a careful manner such as active learning [37-39].

Transfer learning assumes an abundant amount of low-quality data to be present, while only a limited dataset of high-quality data is available. Here, a machine learning model is "pretrained" on the large amount of low-quality data to obtain a model which is generalizable but offers a limited accuracy. Subsequently, the model is refined by training on more high-quality data to improve the accuracy of the predictions. Transfer learning has previously been applied for the prediction of solvation free energies and thermochemical properties [34, 35]. Grambow et al. applied transfer learning on a dataset of $\Delta H^0_{f,298}$, S^0_{298}, and C^0_p for 130.000 molecules determined by low-fidelity quantum chemical calculations. The pretrained model was improved by training on high-quality data for ~10.000 molecules, to obtain an accurate machine learning model [35]. A disadvantage of this methodology is that both an extensive amount of low- and high-fidelity quantum chemical calculations are needed for the model development. Moreover, the approach falls short for the prediction of properties which are experimentally measured instead of computationally, as the number of data required is often unfeasible for experimental campaigns. Another approach to refine low-fidelity quantum chemical calculations is so-called delta learning. Here, the goal is to combine low-fidelity calculations with machine learning to obtain high-fidelity accuracies, where the machine learning model learns to predict the error between the low-fidelity and high-fidelity ab initio calculations [40-42]. Ruth et al. improved CCSD calculations to CCSD(T) accuracy for a wide range of molecules by aid of a graph neural

network [43]. However, the disadvantages of delta learning to transfer learning are twofold. A model trained on transfer learning ideally requires only the molecular structure or simple features, whereas delta learning requires the execution of lower level quantum chemical calculation for the machine learning prediction, which is still cumbersome. Moreover, while transfer learning reduces the amount of high-fidelity data needed for the molecular prediction, this is not necessarily the case for delta learning. With delta learning it is not guaranteed that the relation between the inputs and the error between high-fidelity and low-fidelity calculations is easier to model than the relation between the input and the actual high-fidelity property. However, in general delta-learning results in more accurate predictions than transfer learning as one already starts from low-fidelity calculations.

In this work, a transfer learning methodology for molecular property prediction is presented where group additive models are used for pretraining. Based on the model architecture of chemprop [44], a directed-message passing neural network (D-MPNN) followed by a neural network (NN) is used as a machine learning model. The machine learning model takes SMILES or InChI as an input, making it user-friendly as it does not require low-fidelity calculations or optimized 3D-molecular geometries for model predictions. The presented model is pretrained on a large database of molecular properties (~30.000 molecules) determined by group additivity, to obtain a machine learning model which matches the group additive accuracy. Subsequently, the pretrained machine learning model is refined by using a limited set of high-quality data based on high-accuracy quantum chemistry calculations (150-450 molecules). Group additivity allows to generate a large database of low-fidelity data at almost no computational cost, and for a wide range of desired molecules, making it beneficial for transfer learning applications. The proposed methodology is demonstrated for the prediction of the standard enthalpy of formation, the standard molar entropy, and heat capacity at different

temperatures of a wide range of gas-phase hydrocarbons including radicals and carbenium ions. The application domain of this methodology is not limited to neither thermodynamic properties or hydrocarbons, but these cases highlight the validity and limitations of the proposed approach. Three different machine learning models have been constructed for the property prediction of both regular hydrocarbons, radical hydrocarbons, and carbenium ions respectively.

2. Methods

2.1. Datasets

As the development of the low-fidelity group additive dataset is dependent on the high-fidelity quantum chemical dataset, the high- fidelity quantum chemical dataset will be introduced first.

2.1.1. High-Fidelity Quantum Chemical Dataset

The quantum chemical data consists of thermochemical information of 330 acyclic uncharged hydrocarbons, 442 acyclic radicals, and 162 carbenium ions. All datasets contained alkanes, alkenes, alkynes, alkadienes, alkadiynes and alkenynes with a number of carbon atoms varying between two and eleven (see section 3 Supporting Information). The radicals and carbenium ions included resonance and hyperconjugation stabilized structures. The quantum chemical dataset of carbenium ions contained species with 5- and 6-membered rings among which aromatic rings. The data contained the SMILES, $\Delta H^0_{f,298}$, S^0_{298}, and C^0_p with the heat capacity at seven different temperatures being 300 K, 400 K, 500 K, 600K, 800K, 1000 K, and 1500 K. Part of the data for radicals and regular hydrocarbons has been published in previous work on kinetic modeling [45-49], while the remainder can be found in SI. For carbenium ions the database constructed by Ureel et al. was used [50]. All quantum chemical data in this work was

151 determined at the CBS-QB3 level-of-theory, to ensure the compatibility with the employed
152 GAVs. The $H^0_{f,298}$, S^0_{298}, and C^0_p were determined via ideal gas statistical thermodynamics. For
153 this, the rotation around a single bond was treated as a 1-dimensional hindered rotor. The
154 hindrance potential was calculated via a (semi-)relaxed surface scan at B3LYP/6-31G(d) level
155 of theory. Other internal modes were approximated by harmonic oscillators. All quantum
156 chemical data was obtained via the same methodology, for which we refer to the work of Ureel
157 et al. for further information [50].

158 ### 2.1.2. Low-Fidelity Group Additive Dataset

159 The low-fidelity group additive database contains the group additive predictions of a wide
160 variety of relevant molecules. For every of the three cases (regular hydrocarbons, radicals, and
161 carbenium ions), a separate low-fidelity database was constructed. To ensure a proper
162 pretraining of the machine learning model, it is important that the molecules in the low-fidelity
163 dataset are representative of the high-fidelity dataset molecules. Moreover, the low-fidelity data
164 should include molecules where group additivity falls short (e.g. bigger molecules) that are
165 included in the high-fidelity dataset to facilitate an accurate prediction via the machine learning
166 model. If certain molecular groups are present in the high-fidelity but are absent in the low-
167 fidelity data, the model predictions will likely fall short. One of the advantages in using group
168 additivity is that predictions can be made for any type of molecule for which the graph structure
169 is known, and the required GAVs are determined. Therefore, there is much freedom in
170 determining the type and number of molecules to incorporate in the low-fidelity data. To ensure
171 the representativeness of the low-fidelity dataset with the high-fidelity dataset, all high-fidelity
172 molecules were incorporated in the low-fidelity ones. Additionally, every molecule present in
173 the high-fidelity data was “augmented”, meaning that a random number of operations were

performed on that molecule to obtain a new molecule. These operations existed out of adding a single, double or triple bonded carbon atom, or a branch to the original carbon atom. In this way, the original high-fidelity molecules were modified to obtain new molecules which all contain the original molecule as substructure. Every high-fidelity molecule was augmented an equal number of times to make sure that the low-fidelity data was not biased towards a certain structure. The low-fidelity regular hydrocarbon, radical, and carbenium ion database consisted of circa 33.000 molecules each.

The GAVs and NNIs for $\Delta H^0_{f,298}$, S^0_{298}, and C^0_p of regular hydrocarbons, and radicals determined by Sabbe et al. were used [6, 8]. For carbenium ions the GAVs and NNIs calculated by Ureel et al. were applied [50]. The corrections for the number of optical isomers, and internal and external rotational symmetry were applied to obtain the total entropy. The required symmetry number and number of optical isomers for these corrections were determined automatically with Genesys [51]. When multiple resonance structures were possible, the lowest enthalpy structure was used to determine the thermochemical properties [6]. The group additive property prediction for all low-fidelity molecules was performed using an in-house developed Python script.

A comparison between the distribution of the enthalpy of formation of the high-fidelity and low-fidelity dataset for the regular hydrocarbons is presented in Figure 1. The distribution for the standard molar entropy, and number of carbon atoms for all datasets is provided in the supporting information.

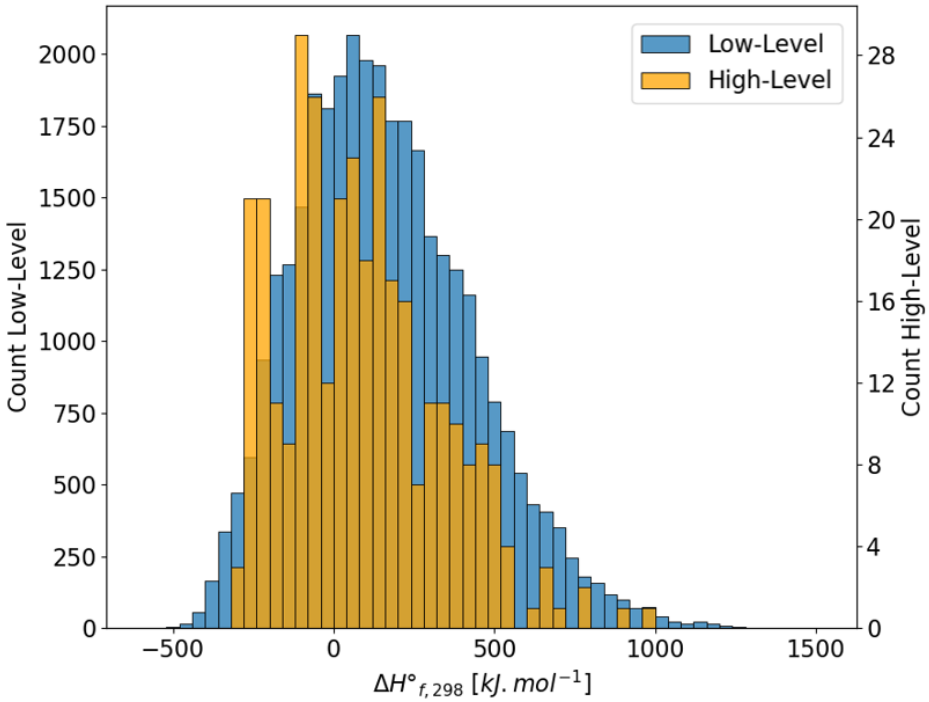

Figure 1. Distribution of the standard enthalpy of formation ($\Delta H^{0}_{f,298}$) for both the high-fidelity and low-fidelity data for regular hydrocarbons.

2.2. Machine Learning Methodology

2.2.1. Machine Learning Model Architecture

The machine learning model consists of a D-MPNN followed by a regular feedforward NN similar to the work of Vermeire et al. [34]. The D-MPNN implementation of chemprop [44] was used in this work, only details relevant to this work and specific adaptations for this work will be discussed. The model takes the SMILES or InChI of the desired molecule as an input which are converted to a graph-based representation via the open-source cheminformatics package RDKit. The molecular properties of the molecule are passed via atom and bond features. The atom features consist of the atomic number, the number of neighboring atoms, the

number of neighboring hydrogen atoms, the atom hybridization, the aromaticity, the atom mass, and the ring size in which the atom is contained. For carbenium ions two additional atom features are added, being the formal charge of the atom and the minimum number of atoms between the charge and the atom. Similarly for radicals the valence of the atom and the minimum numbers of atoms between the unpaired electron and the atom are added as features. The incorporated bond features are the bond type, the conjugation, the ring type, whether the bond is rotatable, and the stereochemistry. For carbenium ions and radicals an additional feature is added to differentiate between bonds neighboring a charged carbon atom or an unpaired electron respectively.

The D-MPNN converts these atom features and bond features to a latent representation of the molecule. To this representation, scaled molecular features are concatenated. These molecular features are the number of aliphatic rings, the number of aromatic rings, the number of rotatable bonds, the molar mass, and a one-hot encoding for the presence of 1,5-interactions and gauche interactions, the global symmetry number, the number of optical isomers, and the natural logarithm of the ratio of the global symmetry number and the number of optical isomers. The last feature is known to be important to determine the total entropy of a molecule [8, 52]. This latent molecular representation serves as an input for the NN which then predicts the nine targets being $\Delta H^0_{f,298}$, S^0_{298}, and C^0_p at 300 K, 400 K, 500 K, 600K, 800K, 1000 K, and 1500 K.

The same model architecture was applied for both the regular hydrocarbons, radicals and carbenium ions. While the model architecture remained the same, a different machine learning model was constructed for every of the three different cases as the thermochemical properties of regular hydrocarbons, radicals and carbenium ions are so different from one another. This would result in unproperly distributed targets leading to a suboptimal performance of the

machine learning model. The hyperparameters of the model were determined by minimizing the root mean squared error (RMSE) on the validation set. The selected hyperparameters were the depth of the D-MPNN, the number of MPNN nodes, the number of NN layers, the number of NN nodes, the number of pretraining epochs, and the number of transfer learning epochs. Around 30 different sets of hyperparameters were investigated for every model type via trial-and-error as this was not within the scope of this work. The employed model architecture consisted of a D-MPNN with a message passing depth of 3 and 600 hidden layers with a LeakyReLu activation function and no dropout or bias. The NN is made up of 3 hidden layers with each a size of 100 nodes with the LeakyReLu activation function, with bias and no dropout. The model optimization is performed via the Adam optimizer [53] and based on the mean squared loss of the normalized targets. The data is split into 10% test data, 10% validation data, and 80% training data. In general, the training data is used for training the model, the validation data for the optimization of hyperparameters and the test data to assess the final model performance. At the end of the training, the model with the lowest loss on the validation set is selected as the final model for evaluation against the separate test set. For every of the three cases, five machine learning models were constructed to allow ensembling after transfer learning, while every of these models were trained on only one fold of the data.

2.2.2. Transfer Learning Methodology

As previously mentioned, three different machine learning models are trained for either the regular hydrocarbons, radicals, and carbenium ions but all are constructed via the same methodology. Initially, the model is pretrained for 700 epochs with a batch size of 10 on the low-fidelity group additivity data to obtain a model which achieves group additive accuracy. Subsequently, the model is refined on the high-fidelity quantum chemical data by training for

30 epochs for the regular hydrocarbons, and carbenium ions, and 40 epochs for radicals. The number of transfer learning epochs was determined by evaluating the decrease in validation error. To avoid any overfitting the number of transfer learning epochs was kept low. During transfer learning, no parameters were frozen. As the radical database is larger than the other two, a longer training procedure is chosen. For the transfer learning, the model parameters of the pretrained model are chosen as initial value and both the parameters of the D-MPNN and NN are refined during the training.

After all hyperparameters were obtained, transfer learning is performed via a standard nested cross-validation or double cross-validation to split the data in a training, validation and test set similar to the work of Dobbelaere et al. [54]. Nested cross-validation consists of an inner and outer loop, where in the outer loop the test data is varied and within every outer loop, the inner loop varies the validation data. Here, 10 folds of the outer loop are performed such that every molecule is exactly one time present in the test data. Next for every of these folds, the validation set is varied in 9 folds such that every remaining molecule is exactly once in the validation set. The nested cross-validation was only employed to evaluate the final model architecture to avoid any bias of the obtained model towards the test data. In this way, the model performance on the entire dataset can be evaluated and there is no bias towards the validation and test data. The model predictions are determined by ensembling the 9 models of the inner loop to obtain a test model prediction of every molecule.

The initialization of the machine learning model before the pretraining is of high-importance due to the large number of parameters in the D-MPNN and feedforward NN. Therefore, an ensemble of five different machine learning models are constructed per case, following the aforementioned procedure. It was observed that the ensemble improved the predictions of the targets compared to a single model.

2.2.3. Model Interpretability

A disadvantage of employing machine learning models compared to group additivity is the loss of interpretability of the constructed models. While with group additivity every structure or group corresponds with a contribution to the predicted property, a machine learning model remains a black-box. To unveil this black-box behind a machine learning model methods such as (SHapley Additive exPlanations) SHAP [55] or (Local interpretable model-agnostic explanations) LIME [56] exist to help interpret these models.

SHAP aims to explain the decisions the model makes by determining the contribution of every input feature to the predicted outcome. For molecular prediction this corresponds to determining the effect of every atom, bond and molecular feature on the predicted outcome. The framework of SHAP is based on Shapley values [57] which originate from game theory and allow to determine the importance of every player to the game result. When using SHAP for machine learning interpretability, these players become the features and the game is the model prediction. SHAP is increasingly being used to interpret various deep learning models and can offer valuable insight in what relations the model exploits [58]. However, the computation of Shapley values is computationally very intensive and scales exponentially $O(2^N)$, therefore several approximative methods exist to determine these values.

Here, the permutation algorithm [59] as implemented in the Python package "shap" is applied to approximate the SHAP-value. The permutation algorithm masks a random permutation of atoms, bonds, and molecular features by a zero-vector to identify the importance of every feature. By sampling the machine learning model output of 10.000 different input permutations of one molecule, a SHAP-value is obtained which provides the effect of every atom, bond, and molecular feature on the observed model output. The sum of all SHAP-values for the features

of one molecule then corresponds to the normalized value of the studied output for that molecule and allows to analyze the effect of every single part of that molecule to the investigated property (Figure 2). In this way, a tip of the black-box machine learning algorithm can be unveiled for the molecular property prediction.

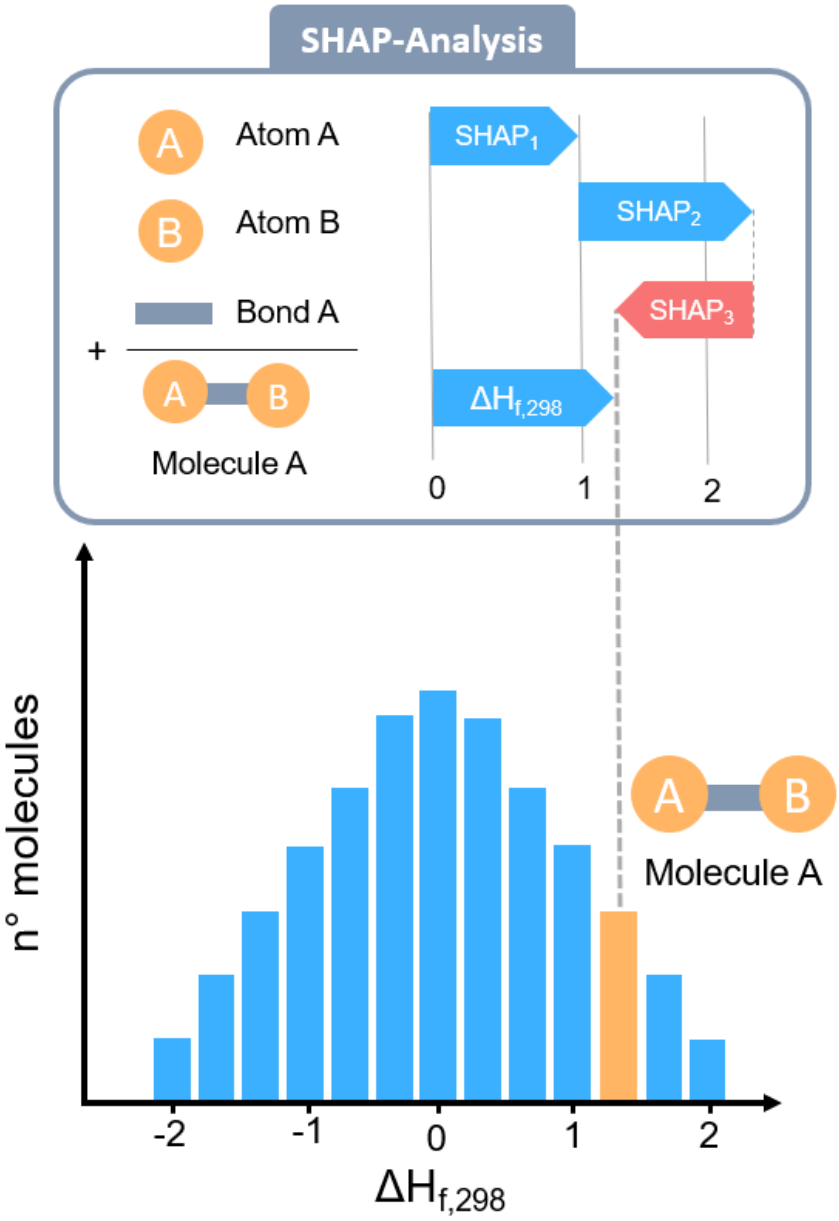

Figure 2. Illustration of SHAP methodology for the interpretation of $\Delta H_{f,298}$ prediction for molecule A, showing the normalized $\Delta H_{f,298}$ distribution of the predictions highlighting the enthalpy value for molecule A. Every atom and bond of the molecule correspond to a SHAP-value which together add up to the normalized $\Delta H_{f,298}$ of the respective molecule. In this case atom B contributes more to the enthalpy than atom A, while bond A results in an enthalpy decrease.

3. Results and Discussion

3.1. Evaluation of Transfer Learning Approach

The presented transfer learning approach is compared with group additivity and a similar machine learning model architecture only trained on high-fidelity ab initio data without pretraining initialization denoted as the non-pretrained model. The model accuracies of these approaches are compared for the regular hydrocarbon dataset in Table 1. The reported errors comprise the predictions for the entire high-fidelity dataset. For the non-pretrained model and transfer learning, the predictions were determined via nested cross-validation while the other models could be directly evaluated on the entire high-fidelity dataset. The prediction accuracy for heat capacities is shown for $C_{p,300}^{0}$ which is representative for the heat capacity at other temperatures. The pretrained model achieves a similar model accuracy as the group additive model, illustrating the adequacy of the pretraining procedure. Moreover, the subsequent transfer learning on the high-fidelity data, improves significantly upon this pretrained model. The transfer learning allows to achieve chemical accuracy in the thermochemical property prediction for a wide variety of hydrocarbons. The importance of the transfer learning procedure

is highlighted by the poor performance of a non-pretrained model. The limited high-fidelity dataset clearly does not allow the model parameters to converge to an optimal value and overfits on the training data.

Table 1. Comparison of the mean absolute error (MAE) and root mean squared error (RMSE) (value between brackets) for the high-fidelity regular hydrocarbon dataset between group additivity (GAV), a non-pretrained D-MPNN + NN model, a pretrained model and a pretrained model refined by transfer learning. The best performing model results are shown in bold.

MAE (RMSE)	$\Delta H^0_{f,298}$ (kJ.mol^{-1})	S^0_{298} (J.mol^{-1}.K^{-1})	$C^0_{p,300}$ (J.mol^{-1}.K^{-1})
GAV	4.44 (5.35)	4.76 (6.83)	2.94 (4.79)
Non-pretrained Model	15.15 (21.61)	10.31 (13.96)	8.30 (11.74)
Pretrained Model	4.57 (5.51)	4.82 (6.87)	2.95 (4.81)
Transfer Learning	**2.52 (3.58)**	**2.97 (4.07)**	**1.66 (2.34)**

Table 2 depicts the four different model accuracies for the radical database for $\Delta H^0_{f,298}$, S^0_{298}, and $C^0_{p,300}$. Long range interactions such as inductive stabilization and steric hindrance significantly influence the stability of radical species. These delocalized effects cannot be incorporated within group additivity resulting in a worse prediction accuracy compared to regular hydrocarbons. Moreover, resonance is more prevalent in radicals which delocalizes the unpaired electron and can pose issues in complex resonance structures. The radical dataset contains many heavy hydrocarbons with multiple double and triple unsaturated bonds resulting in a worse accuracy of the group additive predictions. The pretrained model again matches the accuracy of group additivity. By further refining the pretrained model on the ab initio database, the prediction errors could be reduced four times for $\Delta H^0_{f,298}$ and halved for S^0_{298} and $C^0_{p,300}$

yielding chemically accurate predictions for the wide range of radicals studied. The significant improvement is the result of the radical database being the largest high-fidelity database consisting of 442 radical species of the three studied cases. The non-pretrained model again fails to discover any detailed correlations between the molecular structure and the thermochemical accuracy with an error on the $\Delta H^0_{f,298}$ exceeding 14 kJ.mol^{-1}.

Table 2. Comparison of the mean absolute error (MAE) and root mean squared error (RMSE) (value between brackets) for the high-fidelity radical dataset between group additivity (GAV), a non-pretrained D-MPNN + NN model, a pretrained model and a pretrained model refined by transfer learning. The best performing model results are shown in bold.

MAE (RMSE)	$\Delta H^0_{f,298}$ (kJ.mol^{-1})	S^0_{298} (J.mol^{-1}.K^{-1})	$C^0_{p,300}$ (J.mol^{-1}.K^{-1})
GAV	9.40 (10.80)	5.90 (7.50)	3.25 (4.37)
Non-pretrained Model	14.20 (19.94)	6.49 (8.83)	3.06 (4.21)
Pretrained Model	9.37 (10.76)	5.96 (7.54)	3.24 (4.37)
Transfer Learning	**2.17 (3.30)**	**2.79 (3.85)**	**1.65 (2.31)**

The group additive accuracy for carbenium ions surpasses that of radical species as the GAVs were determined from the same high-fidelity dataset as used in this work. The non-pretrained model performs really poor due to the limited high-fidelity dataset containing data on only 162 carbenium ions. Despite the low amount of data, the transfer learning still allows to improve upon pretrained model for the prediction of the enthalpy. Especially the decrease in RMSE shows that the model improves the worst predictions. However, for entropy and heat capacity transfer learning decreases the model accuracy. Many hypothesis have been evaluated and the suboptimal learning performance is ascribed to the wide variety of carbenium ions including, aromatic and cyclic aliphatic and the low amount of data present, resulting in the machine

learning model failing to detect the trends for entropy and heat capacity. This example again highlights that group additivity, for its staggering simplicity compared to more complex models, remains surprisingly accurate highlighting its suitability for pretraining.

Table 3. Comparison of the mean absolute error (MAE) and root mean squared error (RMSE) (value between brackets) for the high-fidelity carbenium ion dataset between group additivity (GAV), a non-pretrained D-MPNN + NN model, a pretrained model and a pretrained model refined by transfer learning. The best performing model results are shown in bold.

MAE (RMSE)	$\Delta H^0_{f,298}$ (kJ.mol^{-1})	S^0_{298} (J.mol^{-1}.K^{-1})	$C^0_{p,300}$ (J.mol^{-1}.K^{-1})
GAV	4.82 (6.62)	**4.92 (6.51)**	**1.84 (2.57)**
Non-pretrained Model	23.99 (31.19)	13.06 (18.29)	5.33 (7.40)
Pretrained Model	4.86 (6.70)	**4.92 (6.49)**	**1.83 (2.58)**
Transfer Learning	**3.67 (4.92)**	6.04 (7.91)	1.97 (2.83)

Overall, it is clear that the transfer learning approach allows to significantly improve the model accuracy over group additivity with only a limited amount of ab initio data. The predictions for regular hydrocarbons and radical hydrocarbons achieve chemical accuracy for a wide range of species based on less than 450 datapoints. Also for carbenium ions an improvement in the $\Delta H^0_{f,298}$ prediction is found, especially for the worst predictions, based on only 162 ab initio molecules. The improvement of the refined machine learning models over group additivity will be examined in detail for the three separate cases. Based on an analysis with SHAP-values, the discovered machine learning correlations are revealed.

3.2. Transfer Learning for Hydrocarbons

The main limitation of group additive models for regular hydrocarbons is the incorporation of steric effects. The size of a substituent cannot be incorporated in either GAVs or NNIs, which

is evidently an important influence on steric hindrance. Figure 3 (top) depicts the molecules with the largest error on the $\Delta H^0_{f,298}$ for the group additive predictions, up to 16 kJ mol^{-1}. All of these molecules contain branches that experience steric effects from neighboring branches. Molecule (a) and (d) contain branched substituents whereas molecule (b) and (c) have linear substituents. Consequently, group additivity overestimates the stability of molecules (a) and (d) while underestimating molecules (b) and (c) as group additivity considered an "average" steric hindrance in the GAV of the C-(C)$_4$ group. Group additivity cannot differentiate between different substituents as it only looks at the presence of certain substructures resulting in these large discrepancies between the predictions and the ab initio values. For the worst group additive predictions (a-d), the transfer learning model improves only minorly over the group additive predictions. These challenging multi-branched molecules remain exotic and difficult to model compounds. However, overall the MAE on the standard enthalpy of formation for the regular hydrocarbon dataset is reduced from 4.44 kJ.mol^{-1} to 2.52 kJ.mol^{-1} as depicted in Table 1.

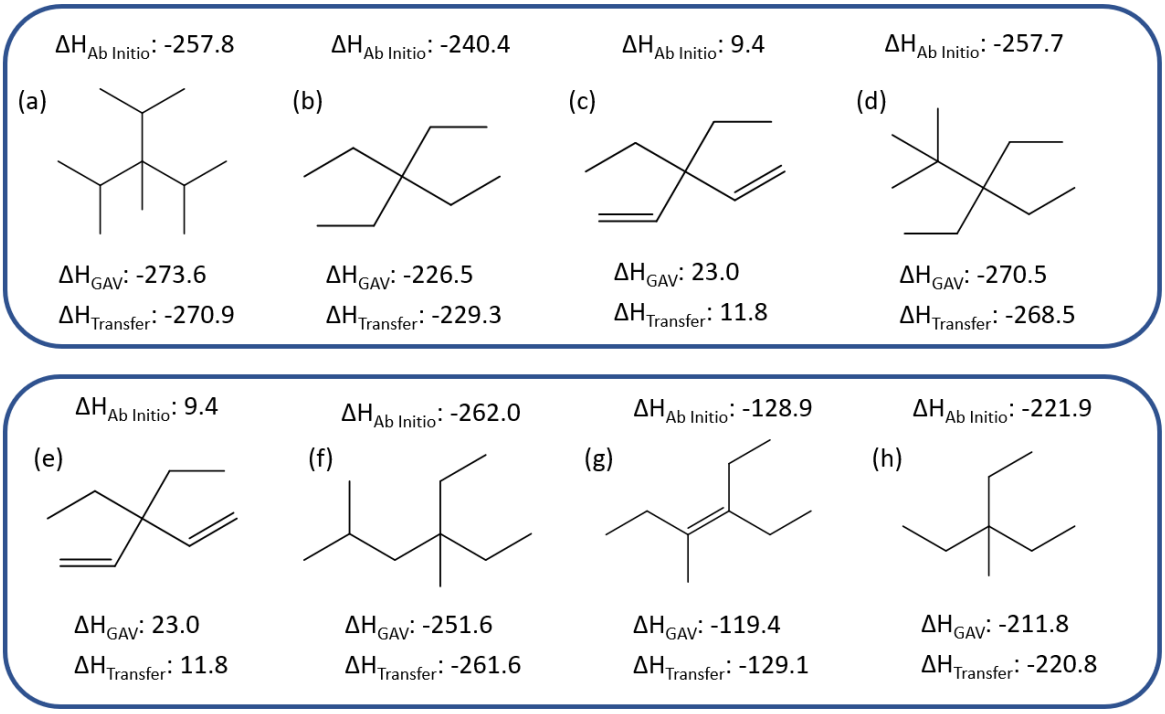

Figure 3. The molecular structures and the predicted and ab initio values of the $\Delta H^0_{f,298}$ (kJ.mol^{-1}) for the worst group additive predictions (top) and the most improved predictions for transfer learning (bottom).

Figure 3 (bottom) illustrates the molecules with the greatest improvement for transfer learning. While the GAV-based predictions fail to incorporate the steric effects, the transfer learning predictions succeeded in this for the four presented molecules. The main improvement of the transfer learning approach is for hydrocarbons with linear substituents which enthalpy has been overestimated by group additivity. This effect is more easily captured by models as the linear chains are less complex in structure than more bulky branched substituents.

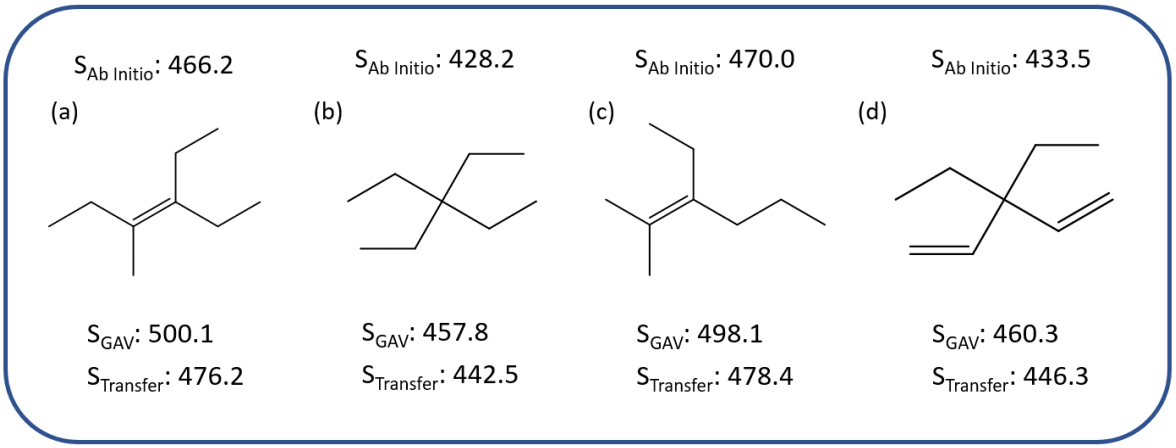

Figure 4. The molecular structures and the predicted and ab initio values of the S^0_{298} (J.mol^{-1}.K^{-1}) for the worst group additive predictions

Figure 4 shows the worst group additive predictions for regular hydrocarbons of the standard molar entropy. The steric effects that limit the rotational freedom of the branches, and hence decrease the entropy, cannot be accounted for in the group additive model, with the GAV-based predictions overestimating the actual entropy. By applying transfer learning, these predictions are significantly improved with the average error for these four complex hydrocarbons decreasing from 29.59 J.mol^{-1}.K^{-1} to 11.34 J.mol^{-1}.K^{-1}.

To be able to shed light onto the machine learning predictions an analysis based on SHAP is performed. The SHAP analysis indicates that the developed models capture the underlying physical principles well. A representative example of the elaborate analysis for hydrocarbons is given in Figure 5. It should be noted that the SHAP-values are an approximation of their true value as the calculation of the exact value is too computationally intensive as specified in the method section. Nevertheless, these values yield valuable information on the model predictions. The presented SHAP-values denote the contribution of every atom to the normalized enthalpy of the molecule, normalized with respect to the prediction distribution. Consequently, a negative SHAP-value does not necessarily denote a negative contribution to the enthalpy but

only a negative contribution to the average enthalpy. In Figure 5 the SHAP-values for the pretrained and transfer learning predictions of 3-ethyl-3-methylpentane are illustrated. Only the SHAP-values of the atoms are depicted, as the SHAP-values for the atoms were found to be an order of magnitude higher than those for bonds. The pretrained model identifies that the C-$(C)_4$ is the most destabilizing for the enthalpy of formation compared to the other groups similar to group additivity. By performing transfer learning, the predicted enthalpy of formation for 3-ethyl-3-methylpentane has decreased. After transfer learning, the machine learning model can differentiate between the three methyl groups on the ethyl chains and the methyl directly bonded to quaternary carbon. The enthalpy decrease is mainly contributed due to an additional stabilizing effect of the end-chain methyl group bonded to the quaternary carbon. clearly indicating the decrease in steric hindrance, as previously discussed. The SHAP-value of the other groups remain mainly unchanged representing the adequacy of group additivity in representing regular hydrocarbons. The analysis based on SHAP clearly indicates that the D-MPNN+NN can excellently discover physical relations based on group additivity and transfer learning.

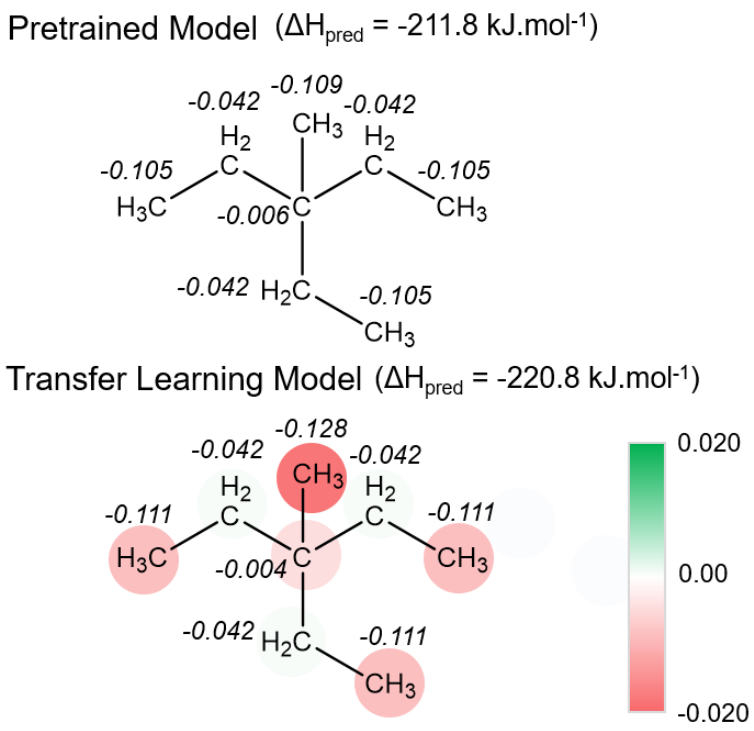

Figure 5. SHAP-values of the atoms for the $\Delta H^0_{f,298}$ prediction of 3-ethyl-3-methylpentane ($\Delta H_{Ab\ Initio}$ = -221.9 kJ.mol^{-1}) for the pretrained model (top) and transfer learning model (bottom) with the corresponding model predictions for $\Delta H^0_{f,298}$. The SHAP-values are shown on top or below the respective carbon atom. The colored dots on the carbon atom depict the change in SHAP-value between the pretrained and transfer learning model.

3.3. Transfer Learning for Radicals

The stability of a radical depends on many non-localized interactions such as resonance stabilization, inductive stabilization and steric hindrance which are challenging to incorporate in group additivity. Figure 6 (top) illustrates the limitations of group additivity by depicting the four worst enthalpy predictions. Radicals (a-c) are all stabilized by a delocalized resonance stabilization which cannot be captured by a single group. Group additivity only considers one

of the resonance structures when estimating the thermodynamic properties, resulting in inaccurate predictions when the resonance structures cannot be represented by one group. Therefore, the group additive predictions significantly overestimate the standard enthalpy of formation of molecule (a-c). Radical (d) is a highly unstable radical ($\Delta H^0_{f,298}$= 660.6 kJ.mol^{-1}) with four different resonance structures of which three secondary radical allenes and one tertiary radical (depicted in (Fig 6.d)). The original group additive value for $C^{\bullet}$-$(C_t)_3$ was determined on the 3-ethynylpenta-1,4-diyn-3-yl which has three primary radical allene resonance structures. As the primary radical allene resonance structures are much more unstable than the corresponding secondary radical group, the enthalpy of radical (d) is overestimated by 27.1 kJ.mol^{-1}. These examples showcase how group additivity fails to incorporate delocalized resonance stabilization.

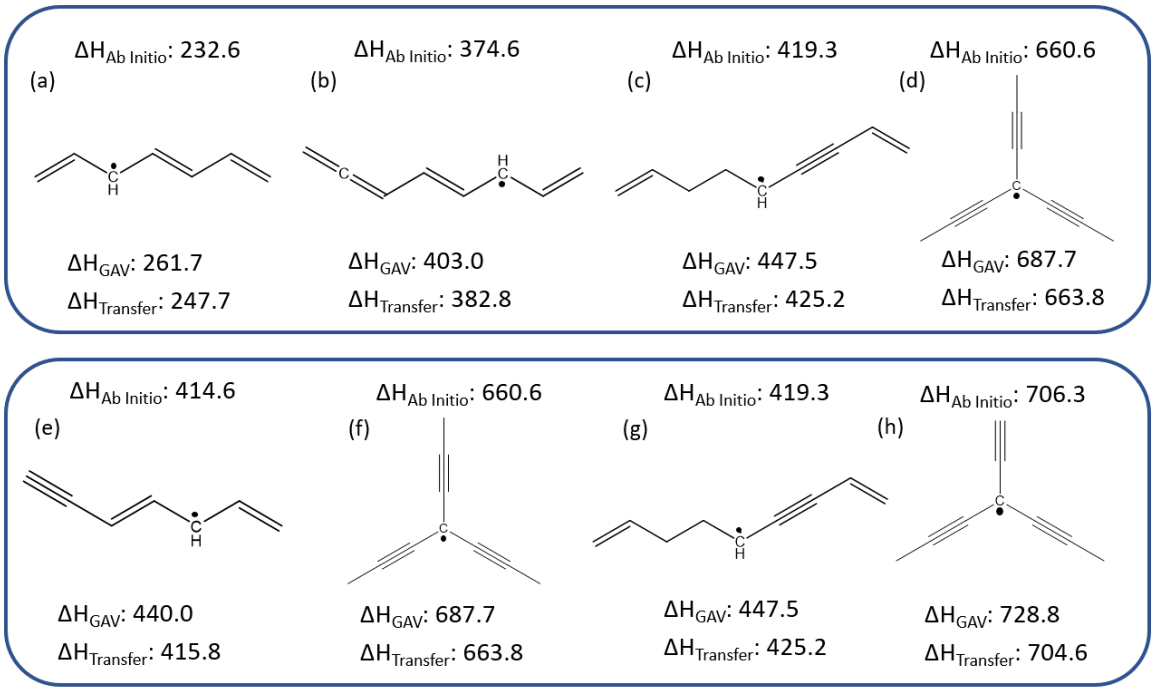

Figure 6. The molecular structures and the predicted and ab initio values of the $\Delta H^0_{f,298}$ (kJ.mol^{-1}) for the worst group additive predictions of radical hydrocarbons (top) and the most improved predictions for transfer learning (bottom).

Figure 6 (bottom) indicates where transfer learning improved the most upon the group additive model. Molecules (e-h) are very similar in nature to molecules (a-d), illustrating the excellent performance of the proposed transfer learning methodology in incorporating these delocalized effects. Of the presented four molecules the mean absolute error for $\Delta H^0_{f,298}$ decreases from 25.8 kJ.mol^{-1} to 3.0 kJ.mol^{-1}. These results illustrate that even when the group additive predictions are distant from the ab initio value, the transfer learning approach is able to correct itself with a limited amount of training data (443 radicals).

Figure 7 shows the worst group additive predictions for radical hydrocarbons of the standard molar entropy. With molecule (a) the steric hindrance of the two cis-interactions is overestimated, resulting in an underestimation of standard molar entropy by group additivity. For molecule (b-d) there is an overestimation of the entropy as the branched molecules do not have a complete rotational freedom. The transfer learning procedure allows to substantially improve upon these entropy predictions for all four of the presented molecules. This is also seen from the overall improvement on the entropy and heat capacity MAE from 5.90 to 2.79 J.mol^{-1}.K^{-1} and 3.25 to 1.65 J.mol^{-1}.K^{-1} respectively as mentioned in Table 2.

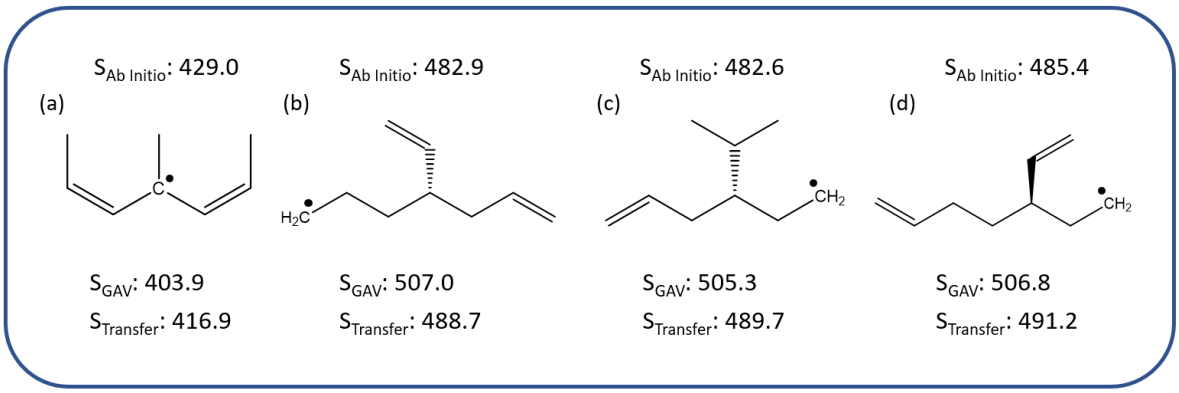

Figure 7. The molecular structures and the predicted and ab initio values of the S^0_{298} (J.mol^{-1}.K^{-1}) for the worst group additive predictions of radical hydrocarbons.

To investigate whether the model really corrects for inductive and resonance stabilization, its predictions were examined with SHAP. As illustrative example, Figure 8 depicts the SHAP analysis for the standard enthalpy of formation of hepta-3,6-dien-1-yn-5-yl for the pretrained model (resembling group additivity) and the refined transfer learning model. The SHAP-values of both atom features, bond features, and molecular features were determined but the values for atom features were found to be the most significant and are presented in Figure 8. In this way, the learning process of the machine learning model can be analyzed and further insights in how the model comes to its predictions are extracted.

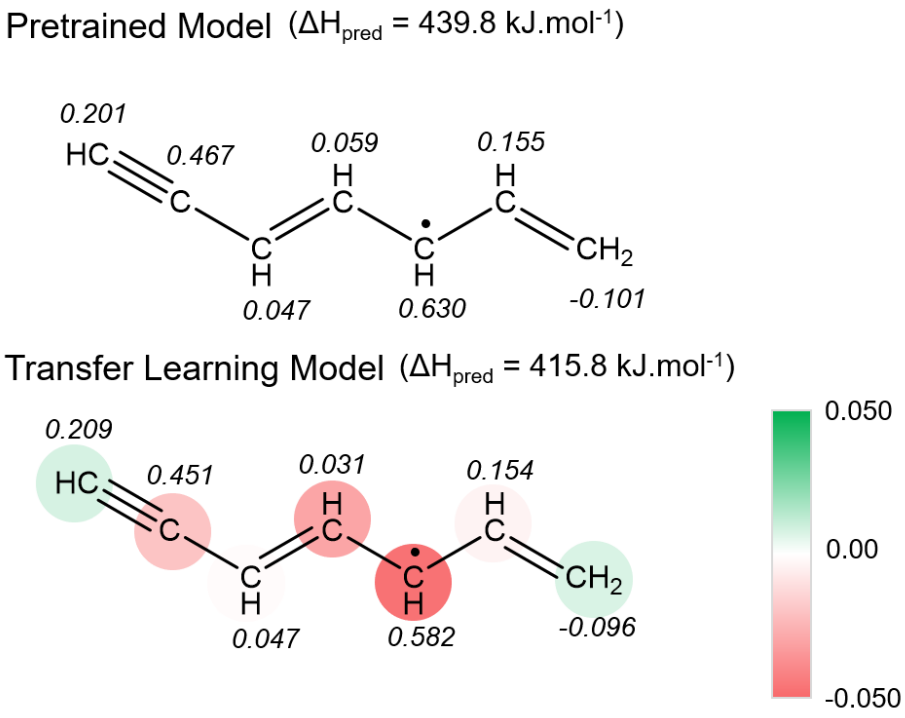

Figure 8. SHAP-values of the atoms for the $\Delta H^0_{f,298}$ prediction of hepta-3,6-dien-1-yn-5-yl ($\Delta H_{Ab\ Initio}$ = 414.6 kJ.mol^{-1}) for the pretrained model (top) and transfer learning model (bottom) with the corresponding model predictions for $\Delta H^0_{f,298}$. The SHAP-values are shown

on top or below the respective carbon atom. The colored dots on the carbon atom depict the change in SHAP-value between the pretrained and transfer learning model.

The SHAP-values clearly illustrate that the machine learning model learns physical trends in both the pretrained as transfer learning model. It observes that the presence of a radical and a triple or double bond will result in a higher enthalpy increase than average and correctly identifies which groups result in the highest enthalpy contributions. Moreover, the change in SHAP-values allows to shed a light onto the transfer learning procedure. The change in SHAP-values is depicted by the colored dots of the respective atoms, with red being a decrease in SHAP while green represents an increase. The transfer learning procedure greatly reduces the enthalpy contribution of the radical group. Hepta-3,6-dien-1-yn-5-yl has three resonance structures, of which only two can be considered by group additivity in the presented representation by the $CH^{\bullet}$-$(C_d)_2$ group. The resonance over the triple bond is not incorporated for the group additive prediction. Via transfer learning this effect is learned by the machine learning model resulting in a decrease in enthalpy contribution of the radical atom. Moreover, the enthalpy contribution of the atoms contributing to this resonance are also decreased by transfer learning. This nicely illustrates how the machine learning model incorporates the delocalized resonance stabilization and is focused on the relevant patterns in the molecules. The obtained SHAP-values indicate that the D-MPNN+NN architecture proposed by chemprop actually follows a physical relation and allows to verify that the model is not overfitted. If the model would be overfitted the SHAP-values would show no trend or be physically explainable. For example, the radical carbon atom might be considered to contribute the most to the molecular stability or the SHAP-values might almost seem random. As the SHAP-values follow physical relations and are not random one can conclude that the model is not overfitted. This

type of analysis however does not allow to learn causal effects but allows to shed light on the general applicability of the model. Moreover, it also proves the adequacy of the presented transfer learning procedure in extracting novel trends starting from a limited number of high-fidelity ab initio data.

3.4. Transfer Learning for Carbenium ions

Group-additive models are excellent for creating interpretable models but fail to incorporate non-linear effects. An example of these non-linear contributions are the previously discussed effects of steric repulsion and inductive stabilization which are prevalent in carbenium ions.

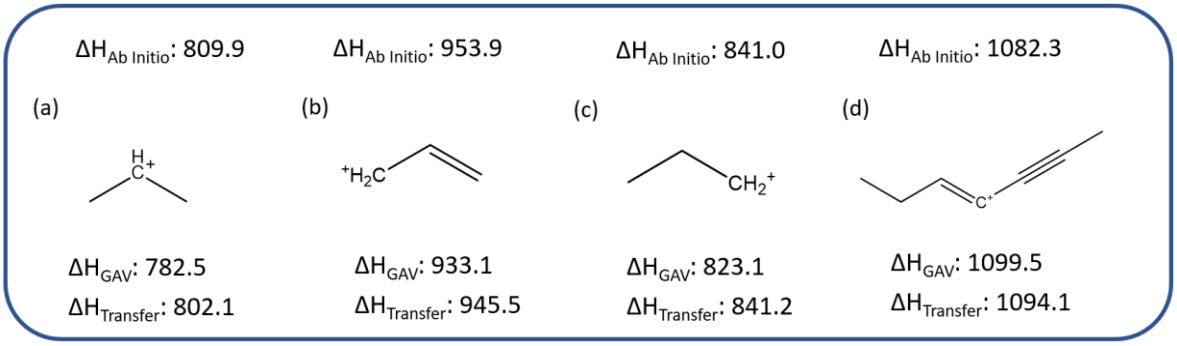

Figure 9. The molecular structures and the predicted and ab initio values of the $\Delta H^{0}_{f,298}$ ($kJ.mol^{-1}$) for the worst group additive predictions of carbenium ions

Figure 9 depicts the worst group additive predictions for carbenium ions of the standard enthalpy of formation. For the small ions (a-c) the previously discussed inductive stabilization is the main source of deviation for group additivity. The transfer learning model can excellently correct the group additive approach to improve upon these predictions. For molecule (d) the group additive model underestimates the stability of the carbenium ion. This is a difficult group due to the resonance stabilization which delocalizes the positive charge resulting in a complex charge density. Due to this, the alkyl chains present are more stabilizing than what is expected

from the graph structure, as these are in closer proximity to a positive charge. While the transfer learning model improves the enthalpy prediction, its predictions are still inaccurate.

The incorporation of steric effects is another limitation of group additivity. The steric effects are prevalent in aromatic carbenium ions as the substituents on the aromatic ring influence each other. Figure 10 depicts the aromatics which are influenced by steric interactions and the group additive and transfer learning predictions on the standard enthalpy of formation. It is clear that the transfer learning does not improve upon the group additive predictions. The steric effects are much more difficult to be captured than inductive stabilization as these effects are dependent on the substituent size. Therefore, steric hindrance is much more difficult to incorporate, definitely because the high-fidelity data in carbenium ions is limited to 162 molecules. The incorporation of steric effects might be improved by providing further information about the 3D-geometry of molecules. Therefore, graph neural networks using 3D-conformers as input could be a valid strategy for future work.

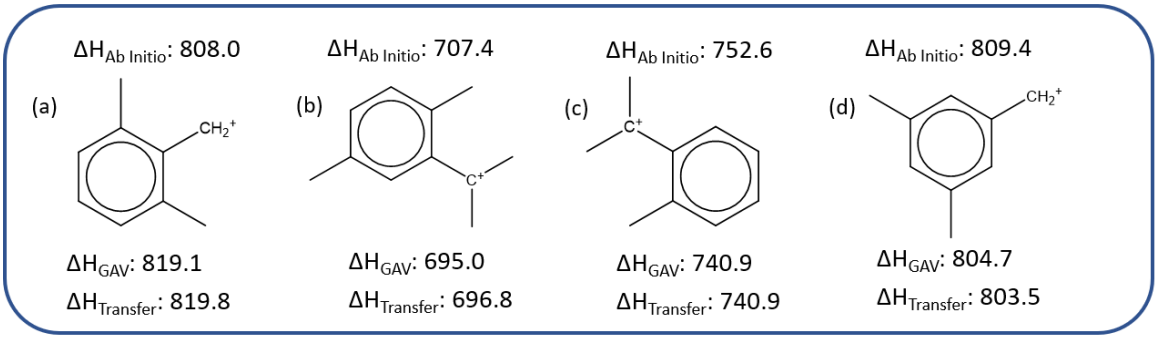

Figure 10. The molecular structures and the predicted and ab initio values of the $\Delta H^{0}_{f,298}$ (kJ.mol^{-1}) for various aromatic carbenium ions exhibiting steric effects.

Figure 11 depicts the error on the group-additive predicted values for the standard enthalpy of formation for two types of carbenium ions with increasing alkyl chain length. There is a clear

bias with group additivity in overestimating the stability of short-chained species and underestimating the stability of long-chained compounds. The deviation between the group-additive values and the ab initio determined values is entirely determined by the inductive stabilization of the alkyl group. The inductive stability is beneficial for the stability of carbenium ions with an enthalpy difference of 41.43 $kJ.mol^{-1}$ between 2-propylium and 2-nonylium due to the inductive effect. However, the inductive effect is delocalized and non-linear and therefore impossible to describe based on group additivity resulting in a discrepancy between ab initio calculations and group additive predictions. Additionally, the inductive effect is structure-dependent as can be deduced from Figure 11. The gained stability depends on the stability of the carbenium-group with a higher instability of the carbenium-group resulting in an increased importance of the inductive effect. For the allylium-group the inductive effect is less pronounced, as the carbenium ion is already resonance stabilized. Therefore, an even larger effect of the inductive stabilization is expected for primary alkyl carbenium ions and carbenium ions neighboring a triple-bonded carbon atom. As the inductive effect is group-dependent no accurate empirical relation can be proposed to account for this effect.

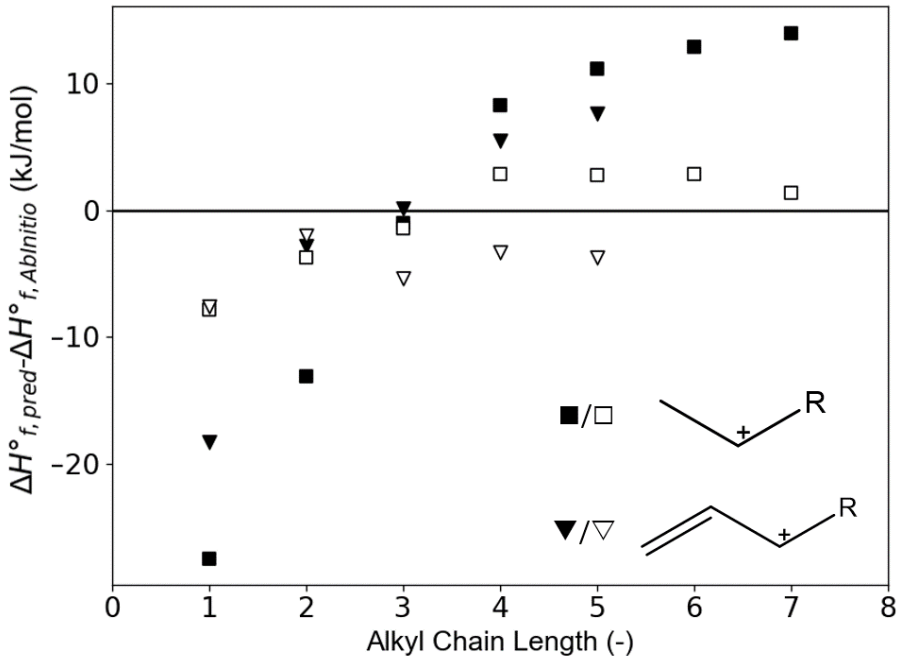

Figure 11. Error between GAV predictions (■/▼), transfer learning predictions (□/▽) and ab initio calculations on standard enthalpy of formation for increasing alkyl chain length of a secondary carbenium ion (■/□) and an allylium-group (▼/▽).

538

539 Overall, the stabilization of alkyl chains neighboring the positively charged carbon atom is
540 significantly underestimated as this value for the GAVs is taken from regular hydrocarbons [50]
541 and considers no inductive stabilization. Hence, the stability of the C^+-$(C)_2(H)$ is overestimated
542 to compensate the lack to incorporate the inductive stabilization effect. Consequently, the
543 largest deviations of the GAV-predicted values of carbenium ions are a result of the failure to
544 include this effect. Therefore, the error on the ΔH^0_{298} prediction is the largest for i-propylium (-
545 27.45 kJ.mol^{-1}), allylium (-20.83 kJ.mol^{-1}), and n-propylium (-18.37 kJ.mol^{-1}) as illustrated in
546 Figure 9. For long-chained molecules this effect is less distinct as these were more present in
547 the training data of the group additive model, due to which our GAVs perform adequately for
548 these molecules.

A more detailed investigation of the nature of the effect of inductive stabilization can be performed based on SHAP. Figure 12 depicts the SHAP-values of the atoms for nonan-2-yl for both the pretrained model trained on group additive predictions (top) and the transfer learning model (bottom). The pretrained model identifies that the positively charged carbon atom gives a more positive effect on the enthalpy of formation. It should be noted that the depicted SHAP-values give the contribution of every atom to the deviation of the predicted enthalpy to the average enthalpy. As the nonan-2-yl is one of the more stable molecules this results in negative enthalpy contributions for every atom which does not mean that they are all stabilizing. The transfer learning predictions clearly improve upon group additivity as illustrated in Figure 11 by the square for alkyl chain length 7. The SHAP-values show that the transfer learning approach is clearly able to correct the group additive model for inductive stabilization. The transfer learning model identifies that the contribution to the enthalpy of formation of the charged carbon group is underestimated by group additivity, and the SHAP-value is increased for the charged carbon. Moreover, it identifies that the stabilizing effect of alkyl chain is underestimated in the pretrained model and corrects the stabilization of these groups.

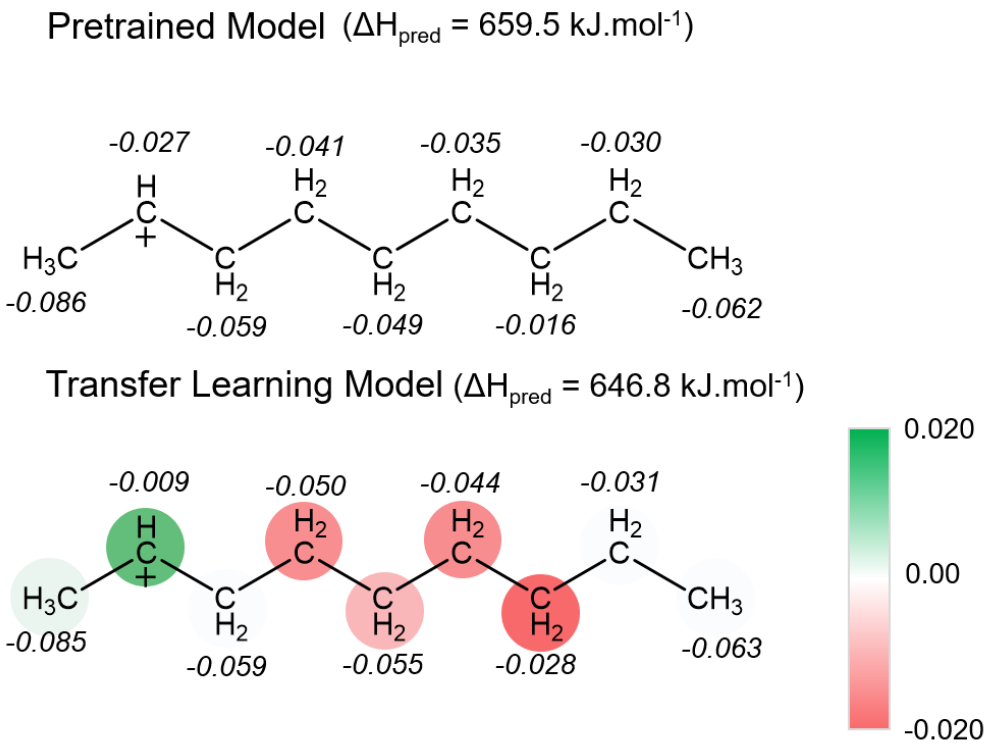

Figure 12. SHAP-values of the atoms for the $\Delta H^0_{f,298}$ prediction of nonan-2-yl ($\Delta H_{Ab\ Initio}$ = 645.4 kJ.mol^{-1}) for the pretrained model (top) and transfer learning model (bottom) with the corresponding model predictions for $\Delta H^0_{f,298}$. The SHAP-values are shown on top or below the respective carbon atom. The colored dots on the carbon atom depict the change in SHAP-value between the pretrained and transfer learning model.

With the transfer learning approach, the machine learning model can clearly improve upon the group additive enthalpy predictions of the molecules with increasing chain length of Figure 11. While the group additive predictions are clearly biased due to the inductive stabilization, this is less the case for the machine learning predictions. Definitely for the larger carbenium ions does the machine learning model take inductive stabilization properly into account.

4. Conclusion

By exploiting group additivity with transfer learning, the power of a highly accurate and data-efficient molecular prediction strategy is illustrated. Accurate thermochemical predictions were obtained for a wide range of hydrocarbons, radicals, and carbenium ions based on 300-450 ab initio calculations for every type. While group additivity is an excellent approach to develop interpretable molecular models with a low amount of data, it only allows to incorporate localized effects. Therefore, transfer learning is employed to improve upon these models and achieve chemical accuracy for complex molecules by including steric effects, inductive stabilization, and delocalized resonance effects, while maintaining the data-efficiency. The transfer learning methodology allowed to halve the error for the hydrocarbon, thermochemical predictions and reduce the prediction error even by a factor four for hydrocarbon radicals. By investigating the obtained machine learning models based on SHAP, it was confirmed that inductive stabilization was excellently incorporated within the models, while steric effects remain more difficult to capture. The utilization of SHAP demonstrates that the D-MPNN+NN, through rigorous pretraining and subsequent refinement via transfer learning, adheres to physical relationships, despite machine learning models typically being regarded as black-box algorithms that lack interpretability when compared to group additivity. In this way, the present approach allows to preserve the benefits of group additivity with a minimal loss in interpretability while significantly improving upon the model accuracy. The successful application of our developed approach in the case study of hydrocarbons, radicals and carbenium ions shows that its application might be promising for the wider field of molecular property or group additive predictions.

Acknowledgements:

We thank Dr. Hans-Heinrich Carstensen for the quantum chemical data on uncharged hydrocarbons and hydrocarbon radicals used in this work. Yannick Ureel acknowledges financial support from the Fund for Scientific Research Flanders (FWO Flanders) through the doctoral fellowship grant 1185822N. The authors acknowledge funding from the European Research Council under the European Union's Horizon 2020 research and innovation programme / ERC grant agreement n° 818607. The computational resources (Stevin Supercomputer Infrastructure) and services used in this work were provided by the VSC (Flemish Supercomputer Center), funded by Ghent University, FWO and the Flemish Government – department EWI.

Supporting Information:

S1. Enthalpy distribution of high- and low-fidelity data. S2. Entropy distribution of high- and low-fidelity data. S3. Distribution of the number of carbon atoms in the high- and low-fidelity data. S4. High-fidelity dataset for regular hydrocarbons (a), radicals (b), and carbenium ions (c). S5. Low-fidelity dataset for regular hydrocarbons (a), radicals (b), and carbenium ions (c). The three trained machine learning models for the property prediction are available on GitHub https://github.com/yureel/HC-Thermochemistry

5. References

[1] Shields BJ, Stevens J, Li J, Parasram M, Damani F, Alvarado JIM, et al. Bayesian reaction optimization as a tool for chemical synthesis. Nature 2021;590(7844):89-96.

[2] Butler KT, Davies DW, Cartwright H, Isayev O, Walsh A. Machine learning for molecular and materials science. *Nature.* 559. Nature Publishing Group; 2018:547-55.

[3] Thybaut JW, Marin GB. Single-Event MicroKinetics: Catalyst design for complex reaction networks. Journal of Catalysis 2013;308:352-62.

[4] Sabbe MK, Reyniers M-F, Reuter K. First-principles kinetic modeling in heterogeneous catalysis: an industrial perspective on best-practice, gaps and needs. Catalysis Science & Technology 2012;2(10):2010-24.

[5] Benson SW, Golden DM, Haugen GR, Shaw R, Cruickshank FR, Rodgers AS, et al. Additivity rules for the estimation of thermochemical properties. Chemical Reviews 1969;69(3):279-324.
[6] Sabbe MK, Saeys M, Reyniers M-F, Marin GB, Van Speybroeck V, Waroquier M. Group additive values for the gas phase standard enthalpy of formation of hydrocarbons and hydrocarbon radicals. The Journal of Physical Chemistry A 2005;109(33):7466-80.
[7] Cohen N. Revised group additivity values for enthalpies of formation (at 298 K) of carbon–hydrogen and carbon–hydrogen–oxygen compounds. Journal of Physical and Chemical Reference Data 1996;25(6):1411-81.
[8] Sabbe MK, De Vleeschouwer F, Reyniers M-F, Waroquier M, Marin GB. First principles based group additive values for the gas phase standard entropy and heat capacity of hydrocarbons and hydrocarbon radicals. The Journal of Physical Chemistry A 2008;112(47):12235-51.
[9] Vandeputte AG, Sabbe MK, Reyniers M-F, Marin GB. Modeling the Gas-Phase Thermochemistry of Organosulfur Compounds. Chemistry – A European Journal 2011;17(27):7656-73.
[10] Paraskevas PD, Sabbe MK, Reyniers MF, Papayannakos N, Marin GB. Group additive values for the gas-phase standard enthalpy of formation, entropy and heat capacity of oxygenates. Chemistry–A European Journal 2013;19(48):16431-52.
[11] Ince A, Carstensen H-H, Sabbe M, Reyniers M-F, Marin GB. Modeling of thermodynamics of substituted toluene derivatives and benzylic radicals via group additivity. AIChE Journal 2018;64(10):3649-61.
[12] Ince A, Carstensen HH, Sabbe M, Reyniers MF, Marin GB. Group additive modeling of substituent effects in monocyclic aromatic hydrocarbon radicals. AIChE Journal 2017;63(6):2089-106.
[13] Naef R, Acree WE. Calculation of the Surface Tension of Ordinary Organic and Ionic Liquids by Means of a Generally Applicable Computer Algorithm Based on the Group-Additivity Method. Molecules 2018;23(5):1224.
[14] Acree Jr W, Chickos JS. Phase transition enthalpy measurements of organic and organometallic compounds and ionic liquids. Sublimation, vaporization, and fusion enthalpies from 1880 to 2015. Part 2. C11–C192. Journal of Physical and Chemical Reference Data 2017;46(1):013104.
[15] Acree Jr W, Chickos JS. Phase transition enthalpy measurements of organic and organometallic compounds. Sublimation, vaporization and fusion enthalpies from 1880 to 2015. Part 1. C1− C10. Journal of Physical and Chemical Reference Data 2016;45(3):033101.
[16] Platts JA, Butina D, Abraham MH, Hersey A. Estimation of Molecular Linear Free Energy Relation Descriptors Using a Group Contribution Approach. Journal of Chemical Information and Computer Sciences 1999;39(5):835-45.
[17] Naef R, Acree WE. Revision and Extension of a Generally Applicable Group-Additivity Method for the Calculation of the Standard Heat of Combustion and Formation of Organic Molecules. Molecules 2021;26(20):6101.
[18] Naef R, Acree Jr WE. Calculation of five thermodynamic molecular descriptors by means of a general computer algorithm based on the group-additivity method: Standard enthalpies of vaporization, sublimation and solvation, and entropy of fusion of ordinary organic molecules and total phase-change entropy of liquid crystals. Molecules 2017;22(7):1059.

[19] Naef R, Acree WE. Calculation of the Vapour Pressure of Organic Molecules by Means of a Group-Additivity Method and Their Resultant Gibbs Free Energy and Entropy of Vaporization at 298.15 K. Molecules 2021;26(4):1045.
[20] Sumathi R, Carstensen HH, Green WH. Reaction rate prediction via group additivity, part 2: H-abstraction from alkenes, alkynes, alcohols, aldehydes, and acids by H atoms. The Journal of Physical Chemistry A 2001;105(39):8969-84.
[21] Sumathi R, Carstensen HH, Green WH. Reaction rate prediction via group additivity part 1: H abstraction from alkanes by H and CH3. The Journal of Physical Chemistry A 2001;105(28):6910-25.
[22] Van de Vijver R, Sabbe MK, Reyniers M-F, Van Geem KM, Marin GB. Ab initio derived group additivity model for intramolecular hydrogen abstraction reactions. Physical Chemistry Chemical Physics 2018;20(16):10877-94.
[23] Sabbe MK, Reyniers MF, Waroquier M, Marin GB. Hydrogen radical additions to unsaturated hydrocarbons and the reverse β-scission reactions: modeling of activation energies and pre-exponential factors. ChemPhysChem 2010;11(1):195-210.
[24] Sabbe MK, Reyniers MF, Van Speybroeck V, Waroquier M, Marin GB. Carbon-centered radical addition and β-scission reactions: modeling of activation energies and pre-exponential factors. ChemPhysChem 2008;9(1):124-40.
[25] Paraskevas PD, Sabbe MK, Reyniers M-F, Papayannakos NG, Marin GB. Group additive kinetics for hydrogen transfer between oxygenates. The Journal of Physical Chemistry A 2015;119(27):6961-80.
[26] Naef R, Acree WE. Application of a General Computer Algorithm Based on the Group-Additivity Method for the Calculation of Two Molecular Descriptors at Both Ends of Dilution: Liquid Viscosity and Activity Coefficient in Water at Infinite Dilution. *Molecules.* 23. 2018.
[27] Conte E, Martinho A, Matos HA, Gani R. Combined Group-Contribution and Atom Connectivity Index-Based Methods for Estimation of Surface Tension and Viscosity. Industrial & Engineering Chemistry Research 2008;47(20):7940-54.
[28] Barrientos EJ, Lapuerta M, Boehman AL. Group additivity in soot formation for the example of C-5 oxygenated hydrocarbon fuels. Combustion and Flame 2013;160(8):1484-98.
[29] Chung Y, Vermeire FH, Wu H, Walker PJ, Abraham MH, Green WH. 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):433-46.
[30] Simamora P, Yalkowsky SH. Group contribution methods for predicting the melting points and boiling points of aromatic compounds. Industrial & engineering chemistry research 1994;33(5):1405-9.
[31] Li Y, Li P, Yang X, Hsieh C-Y, Zhang S, Wang X, et al. Introducing block design in graph neural networks for molecular properties prediction. Chemical Engineering Journal 2021;414:128817.
[32] Aouichaoui ARN, Fan F, Mansouri SS, Abildskov J, Sin G. Combining Group-Contribution Concept and Graph Neural Networks Toward Interpretable Molecular Property Models. Journal of Chemical Information and Modeling 2023;63(3):725-44.
[33] Weiss K, Khoshgoftaar TM, Wang DD. A survey of transfer learning. Journal of Big Data 2016;3(1):9-.

[34] Vermeire FH, Green WH. Transfer learning for solvation free energies: From quantum chemistry to experiments. Chemical Engineering Journal 2021;418:129307-.
[35] Grambow CA, Li Y-P, Green WH. Accurate Thermochemistry with Small Data Sets: A Bond Additivity Correction and Transfer Learning Approach. The Journal of Physical Chemistry A 2019;123(27):5826-35.
[36] Zhong S, Hu J, Yu X, Zhang H. Molecular image-convolutional neural network (CNN) assisted QSAR models for predicting contaminant reactivity toward OH radicals: Transfer learning, data augmentation and model interpretation. Chemical Engineering Journal 2021;408:127998.
[37] Alshehri AS, You F. Deep learning to catalyze inverse molecular design. Chemical Engineering Journal 2022;444:136669.
[38] Settles B. Active learning. Synthesis Lectures on Artificial Intelligence and Machine Learning 2012;18:1-111.
[39] Ureel Y, Dobbelaere MR, Ouyang Y, De Ras K, Sabbe MK, Marin GB, et al. Active Machine Learning for Chemical Engineers: a Bright Future Lies Ahead! 2022.
[40] Bogojeski M, Vogt-Maranto L, Tuckerman ME, Müller K-R, Burke K. Quantum chemical accuracy from density functional approximations via machine learning. Nature Communications 2020;11(1):5223.
[41] Huang B, von Lilienfeld OA. Ab Initio Machine Learning in Chemical Compound Space. Chemical Reviews 2021;121(16):10001-36.
[42] Plehiers PP, Lengyel I, West DH, Marin GB, Stevens CV, Van Geem KM. Fast estimation of standard enthalpy of formation with chemical accuracy by artificial neural network correction of low-level-of-theory ab initio calculations. Chemical Engineering Journal 2021;426:131304.
[43] Ruth M, Gerbig D, Schreiner PR. Machine Learning of Coupled Cluster (T)-Energy Corrections via Delta (Δ)-Learning. Journal of Chemical Theory and Computation 2022;18(8):4846-55.
[44] Yang K, Swanson K, Jin W, Coley C, Eiden P, Gao H, et al. Analyzing Learned Molecular Representations for Property Prediction. Journal of Chemical Information and Modeling 2019;59(8):3370-88.
[45] Xu C, Al Shoaibi AS, Wang C, Carstensen H-H, Dean AM. Kinetic Modeling of Ethane Pyrolysis at High Conversion. The Journal of Physical Chemistry A 2011;115(38):10470-90.
[46] Vervust AJ, Djokic MR, Merchant SS, Carstensen H-H, Long AE, Marin GB, et al. Detailed Experimental and Kinetic Modeling Study of Cyclopentadiene Pyrolysis in the Presence of Ethene. Energy & Fuels 2018;32(3):3920-34.
[47] Vermeire FH, De Bruycker R, Herbinet O, Carstensen H-H, Battin-Leclerc F, Marin GB, et al. Experimental and kinetic modeling study of the pyrolysis and oxidation of 1,5-hexadiene: The reactivity of allylic radicals and their role in the formation of aromatics. Fuel 2017;208:779-90.
[48] Khandavilli MV, Djokic M, Vermeire FH, Carstensen H-H, Van Geem KM, Marin GB. Experimental and Kinetic Modeling Study of Cyclohexane Pyrolysis. Energy & Fuels 2018;32(6):7153-68.
[49] Khandavilli MV, Vermeire FH, Van de Vijver R, Djokic M, Carstensen H-H, Van Geem KM, et al. Group additive modeling of cyclopentane pyrolysis. Journal of Analytical and Applied Pyrolysis 2017;128:437-50.

[50] Ureel Y, Vermeire FH, Sabbe MK, Van Geem KM. Ab Initio Group Additive Values for Thermodynamic Carbenium Ion Property Prediction. Industrial & Engineering Chemistry Research 2022.
[51] Vandewiele NM, Van Geem KM, Reyniers M-F, Marin GB. Genesys: Kinetic model construction using chemo-informatics. Chemical Engineering Journal 2012;207:526-38.
[52] Van Speybroeck V, Vansteenkiste P, Van Neck D, Waroquier M. Why does the uncoupled hindered rotor model work well for the thermodynamics of n-alkanes? Chemical physics letters 2005;402(4-6):479-84.
[53] Kingma DP, Ba J. Adam: A method for stochastic optimization. arXiv preprint arXiv:14126980 2014.
[54] Dobbelaere MR, Ureel Y, Vermeire FH, Tomme L, Stevens CV, Van Geem KM. Machine Learning for Physicochemical Property Prediction of Complex Hydrocarbon Mixtures. Industrial & Engineering Chemistry Research 2022;61(24):8581-94.
[55] Lundberg SM, Lee S-I. A unified approach to interpreting model predictions. Advances in neural information processing systems 2017;30.
[56] Ribeiro MT, Singh S, Guestrin C. " Why should i trust you?" Explaining the predictions of any classifier. *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining.* 2016:1135-44.
[57] Shapley L. Quota solutions op n-person games1. Edited by Emil Artin and Marston Morse 1953:343.
[58] Rozemberczki B, Watson L, Bayer P, Yang H-T, Kiss O, Nilsson S, et al. The Shapley Value in Machine Learning. arXiv preprint arXiv:220205594 2022.
[59] Štrumbelj E, Kononenko I. Explaining prediction models and individual predictions with feature contributions. Knowledge and information systems 2014;41(3):647-65.