A predictive model for the aggregation of polycyclic aromatic compounds

Jacob C. Saldinger[a], Paolo Elvati[b], Angela Violi[a,b,c,*]

[a]*Department of Chemical Engineering, University of Michigan, Ann Arbor, MI, 48109-2136, United States*
[b]*Department of Mechanical Engineering, University of Michigan, Ann Arbor, MI 48109-2125, United States*
[c]*Department of Biophysics, University of Michigan, Ann Arbor, MI 48109-1055, United States*

Abstract

The physical aggregation of polycyclic aromatic compounds (PACs) is a key step in soot inception. In this work, we set out to elucidate which molecular properties influence the physical growth process and use machine learning to quantitatively relate these features to the propensity of these molecules to physically dimerize with other PACs. To this end, we first develop a dataset of PAC monomers along with their calculated free energies of dimerization emphasizing a set of PACs with a diverse range of properties. First, we augment existing calculations of dimerization energies with our own molecular dynamics simulations enhanced by well-tempered Metadyanmics. We then demonstrate that a machine learning model based on the least absolute shrinkage and selection operator (Lasso) is able to quantitatively learn how molecular features contribute to physical aggregation and predict the free energy of dimerization for new pairs of molecules. The model is able to accurately determine the stability for both homodimerization and heterodimerization cases. Our approach also provides a data driven method to determine the molecular features most important to predicting the dimer stability. From this, we determine that the PAC properties most influential to physical dimerization are size, shape, oxygenation, and presence of rotatable bonds. This work highlights the molecular complexity of the PAC monomers that must be accounted for in order to accurately represent physical aggregation. We anticipate that this approach will allow for more effective modeling of the PAC dimerization process as it facilitates the efficient prediction of dimerization propensity from easily calculable molecular features.

Keywords: Molecular Dynamics; Machine Learning; Nucleation; Soot

*Corresponding author: avioli@umich.edu (Angela Violi)
URL: avioli@umich.edu (Angela Violi)

1. Introduction

Central to modeling soot formation in combustion environments is understanding how polycyclic aromatic compounds (PACs) transition into larger nanoparticles. These multi-ringed aromatic structures are believed to transform into soot through chemical and physical pathways [1]. Unpaired electrons on PACs have been observed to react with other radical species to form three dimensional structures [2, 3], while PACs can stack into larger clusters held together by electrostatic and dispersion forces [4, 5, 6]. This aggregation process is believed to be an important component of soot formation as it provides the initial nucleation step or can hold PACs in proximity with each other so other chemical growth mechanisms can occur [7, 8, 9]. For this reason, a proper understanding of the physical inception process is a necessary step towards creating a comprehensive soot model.

A number of experimental and computational studies have sought to characterize how the process of physical aggregation occurs [10, 11]. Miller developed a model that showed the importance of mass in hydrocarbon aggregation and determined that only PACs larger than $800\,\mathrm{Da}$ would exist long enough to play a significant role in physical growth [4]. Other studies have concluded that aggregation can occur at lower masses [12]. While many of these earlier studies were based on PACs within the Stein-Fahr stabilomer grid [13], more recent studies of PAC formation have suggested that these molecules occupy a much more diverse chemical space with properties such as oxygenation, aliphatic branching, and five-membered rings [14, 15, 16, 17, 18, 19]. A number of works have assessed the effects of these properties on the propensity of these molecules to form dimers. Molecular dynamics studies have found that physical dimerization is promoted by aliphatic chains and thus mass alone is not a sufficient descriptor of the process [20, 21]. Moreover, the presence of oxygen [8, 22] and molecular curvature [23] have been shown to affect how these molecules dimerize.

In addition to characterizing the properties that promote dimerization, a number of works have looked to use these properties to make quantitative predictions about the tendencies of molecules to aggregate [6, 24]. As the size of the PAC is understood to be an important property, Herdman and Miller developed a linear relationship between the reduced mass of PAC monomers and their propensity to dimerize [5]. Raj *et al.* showed that the collision efficiency is an important factor in representing dimerization and can be predicted from the mass and shape of constituent PAC molecules [25]. A predictive model for dimer stability has also been developed by fitting the free energy (FE) of aggregation to molecular properties, such as size, number of carbons, and solvent accessible surface area [26].

Although these studies have shown some predictive capacity, they are unable to account for the diverse PAC feature space that has been observed both experimentally [17] and computationally [27] in flames. Recently, Elvati *et al.* examined a number of these properties including size, oxygenation, radius of gyration, and presence of rotatable bonds, finding that all these features affect the aggregation FE land-

scape [28]. This result suggests that models that do not account for these properties are incomplete and highlights the need for a prediction scheme that can identify the complex relationships these molecular properties have on the physical growth process.

To this end, here we introduce a machine learning method for the quantitative prediction of the physical dimerization propensities of PACs. We started by expanding existing data [28] using molecular dynamics simulations, in order to have FE profiles (as a function of molecular distance) for PACs with different functional groups. We performed 141 additional molecular dynamics simulations to create a final dataset containing 105 unique PAC pairs with molecular properties such as aliphatic chains, five-membered rings, oxygenated groups, and aliphatic linkages. We then trained a supervised machine learning model (Lasso method [29]) in order to both predict the FE of aggregation and to select a small set of molecular properties that are key for the prediction. The results underscore how machine learning can be used to process the large feature hyperspace that is associated with more realistic and complex PAC properties in order to create more accurate aggregation and, in the future, kinetic models.

2. Methodology

2.1. Molecular Dynamics

To generate data for the machine learning model, we used the FE profiles of all the possible combinations of 14 PACs (see Fig. S1 in the Supplementary Materials for the structures). This class of molecules contains a diverse set chemical features observed in both experimental [17] and computational [27] studies and is an extension of the one used in previous works [8, 22, 28].

When available, we used previously computed FE profiles [28]; otherwise the missing FE profiles were obtained with the same procedure, briefly reported below. FE profiles were reconstructed by using the well-tempered Metadynamics technique; all simulations were carried out in the NAMD program [30] using the PLUMED plugin [31]. Each simulation consisted of a $1\,\mathrm{ns}$ minimization and equilibration starting from two PAC molecules spaced $1\,\mathrm{nm}$ apart. This step was followed by a $100\,\mathrm{ns}$ simulation at $1000\,\mathrm{K}$ biased on the distance between the molecules' center-of-mass (COM)). The curves of three independent runs were then used to compute the aggregation FE as the difference between the ensemble average for the monomer (0.35-$0.75\,\mathrm{nm}$) and dimer state (3.75-$4.0\,\mathrm{nm}$). Of note, positive values indicate that the molecules are more likely to be found not aggregated, while for negative values the aggregate state is preferred.

2.2. Machine Learning

For each PAC, we computed 312 molecular features describing size, shape, composition, and chemistry with an in-house code. Many of these features have previously

been employed in other quantitative structure-property relationship studies [32, 33] and a detailed list of features is given in the supporting information (see Tab. S1). Since each FE of aggregation depends on two molecules, we combine individual molecular features by computing different types of averages, as discussed in the Results section.

Before training our machine learning model, we eliminated similar features by removing any feature with a variance of zero and any feature with a Pearson correlation greater than 0.95. To build a predictive model for the FE of aggregation, we applied the Lasso method (Scikit-learn implementation [34]), as it allows for high accuracy and often interpretable predictions [29]. Lasso, which uses the least absolute shrinkage and selection operator, has been applied successfully to make interpretable predictions in chemical problems [35] as it eliminates extraneous features and only selects a smaller subset of properties needed to make the predictions. Specifically, it is a supervised machine learning regression model that minimizes a loss function given by:

$$L(\beta, \lambda) = \sum_{i=1}^{n} (Y_i - \beta X_i) + \lambda \sum_{j=1}^{p} \beta_j \tag{1}$$

In this equation, Y is the true value, X are the input features, β is a set of feature weights learned by the model, and λ is a regularization parameter set manually. In other words, Lasso minimizes the sum of the squares of the residuals with a regularization term proportional to the l_1 norm, creating a penalty on feature weights, which results in a more concise model. The further regularization of the l_0 norm, which would produce the simplest model by finding the smallest subset of features that fits the data, is the ideal next step but it is computationally intractable. The l_1 norm employed by Lasso provides an approximation that can be efficiently solved as a convex optimization problem [35].

To avoid artifacts due to the different magnitudes of features affecting our results, training data was first scaled using a standard scaler fit [34] that centers each feature at its mean value and normalizes by the standard deviation. We then used the training data to select the optimal parameter λ (see hyperparameter optimization in Supporting Material) and to train a Lasso model. We tested the model using leave-one-out cross validation (*i.e.*, withholding one FE for each fold and using the remaining as training data) and considered the selected features as those with non-zero coefficients.

3. Results and Discussion

3.1. Free Energy Prediction Model

Our predictive model (Fig. 1) performs well, with a mean absolute error (MAE) of $6.4\,\mathrm{kJ\,mol^{-1}}$, only slightly higher than the average uncertainty of our MD simulations ($3.5\,\mathrm{kJ\,mol^{-1}}$) and the average energy for one translational degree of freedom

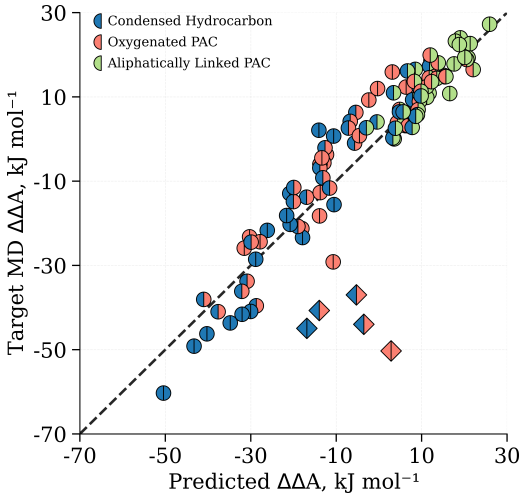

Fig. 1: Comparison between calculated and predicted FE of aggregation at $1000\,\mathrm{K}$. Dashed line provides reference of correct predictions. Color represents dimer component type: green is an aliphatically linked PAC, red is an oxygenated PAC, and blue is a condensed hydrocarbon. Points with two colors share all the corresponding characteristics. Diamonds represent dimer pairs with errors at least twice the RMSE ($10.2\,\mathrm{kJ\,mol^{-1}}$).

($4.184\,\mathrm{kJ\,mol^{-1}}$) at $1000\,\mathrm{K}$ (see "This work" in Fig. 2). To test for information leakage, or in other words that the model is learning from general molecular properties and motifs and not the presence of the same monomer in the training set, we repeated the leave-one-out cross validation while omitting from the training data all samples which share a monomer with the testing sample. As expected, since we are training with less data ($\sim$13% for homo-aggregation and $\sim$26% for hetero-aggregation, smaller dataset), the prediction slightly worsens ("No leakage" in Fig. 2). However, with a root mean squared error (RMSE) of $11.3\,\mathrm{kJ\,mol^{-1}}$ and MAE of $7.6\,\mathrm{kJ\,mol^{-1}}$, the model still performs better than existing models (see Fig. 2).

Interestingly, the RMSE of our predictions is higher than the MAE which suggests that a few interactions are not predicted as well as the rest of the data. The analysis of the data highlights that there are 5 pairs (*i.e.*, AD, AJ, BD, CF, and DE) for which the error in the predicted FE value is twice the RMSE value. For all of them, the predicted aggregate is less stable than the MD simulations would indicate, and four of them involve molecule C or D, which are the only ones in the dataset that have both aliphatic chains and oxygenated groups. We have observed that oxygenation destabilizes the physical aggregations of PACs [8, 22] while aliphatic chains show the opposite trend [20, 21], and when multiple competing features of similar magnitude affect the FE, the outcome is not easy to predict [28]. Moreover, since our dataset lacks molecules that have only aliphatic chains or more cases of similar molecules, it is possible that the model is not able to properly learn the interplay of these particular features.

Our model outperforms existing physical dimerization models presented in the lit-

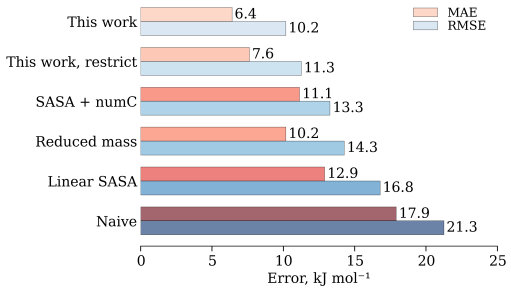

Fig. 2: Comparison between our predictive model and published ones applied to our dataset at $1000\,\mathrm{K}$. Red shows the MAE and blue the RMSE. For reference, the average standard error of the input data (MD simulations) is $\sim 3.5\,\mathrm{kJ\,mol^{-1}}$.

erature, as shown in Fig. 2. We compared our results, including the test on a restricted dataset with no data leakage illustrated above (labeled "This work" and "This work, restrict", respectively) with three additional models. For all methods, we performed a leave-one-out validation procedure, fitting each model's parameters and functions to the molecules in our dataset. First, we compared our results with the widely used model introduced by Herdman and Miller [5], which assumes a linear correlation between the reduced mass and the binding energy (labeled "Reduced mass"). More recently, Lowe *et al.* [26] characterized a number of polycyclic aromatic dimers and developed a predictive model for the change in FE between the monomer and dimer states based around the solvent accessible surface area and number of carbons. For our second comparison, we used the linear fit from the original publication that relates the the average carbon surface area and the FE (labeled linear SASA). Next, instead of using the published linear fit, we instead used the molecular descriptors (number of carbons and solvent accessible surface area) as input features into a Lasso model (labeled as SASA + numC). Finally, we consider the naive case in which all values were predicted as the average free energy of the dataset. In all cases, the predictive model presented in this work performs better than the previous models, showing that a more comprehensive feature set, such as the one employed here, can better capture molecular properties responsible for dimerization.

3.2. Molecular Features Selection

One of the advantages of the selected method, is its ability to provide a degree of interpretability towards the aspects that control the prediction, as it sets coefficients of unused features to zero [29]. Thus, by analyzing which features the Lasso model retains, we can gain a sense of which molecular properties are important for predicting the FE of dimer aggregation. Overall, across all 105 folds of cross validations, the model selects a nearly identical set of 10 features (a complete list of features selected and the number of folds in which they are retained is provided in Tab. S2 of the Supplementary Material). If we exclude these top features, no other feature is selected in more than four folds. Broadly, these features can be divided into three groups

of properties that are important for PAC dimerization: size, shape, and presence of specific chemical groups.

3.2.1. Size

The first class of properties are extrinsic properties that are broadly related to the size of the molecule. Specifically, the algorithm selected the number of aromatic rings, the number of carbons not connected to a hydrogen, the number of tessellations containing four carbons, the number of tessellations with three carbons and a hydrogen, and the number of six-membered rings.

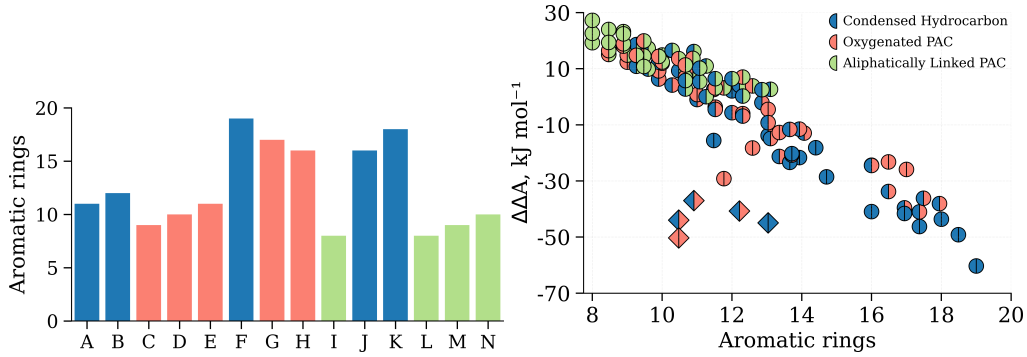

Fig. 3: Relationship between number of aromatic rings and dimerization FE. Left: The number of aromatic rings associated with each dimer. Right: Aggregation propensity compared to average number of aromatic rings in the dimer. The five outliers discussed in the previous section are denoted as diamonds. Colors represent the dimer's component type: green indicates an aliphatically linked PAC, red an oxygenated PAC, and blue a condensed hydrocarbon. Points with two colors share all the corresponding characteristics.

Figure 3 shows that the FE of dimerization is strongly related to the (harmonic) average number of aromatic rings in the dimer (Pearson coefficient of -0.8397 and Spearman coefficient of 0.8719). This result agrees with the general observation that PACs will often cluster in lateral stacks and the interaction strength between PACs is closely related to their number of aromatic rings [36, 37]. Moreover, at least for soot precursors, the number of aromatic rings is closely correlated with mass, hence the use of the latter as a descriptor for the aggregation strength in other works [5].

Among the molecular descriptors in this class, the number of aromatic rings is the feature that has the highest correlation with the FE (more than the number of six membered rings, for example), but it is crucial to note that, by itself, it is not sufficient to fully capture the physical dimerization. A linear fit of the FE as function of the total number of aromatic rings produces predictions model with an RMSE of $15.6\,\mathrm{kJ\,mol^{-1}}$ and a MAE of $11.3\,\mathrm{kJ\,mol^{-1}}$, which has a significantly larger error than our model and is (not coincidentally) comparable to using only the mass as a descriptor (see Fig. 2)

Some of the features in this group encode size with molecular shape information. One such example is the number of internal carbon atoms, defined as the aromatic

carbon atoms that are not bonded to H atoms. As, most of the molecules in the dataset are highly pericondensed hydrocarbons, these PACs will have a greater percentage of internal carbons than catacondensed PACs.

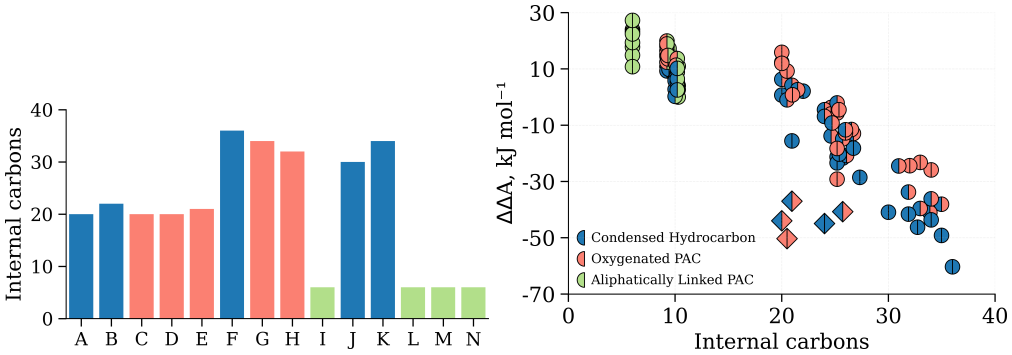

Fig. 4: Relationship between number of internal carbons and dimerization FE. Left: The number of internal carbons associated with each monomer. Right: Aggregation propensity compared to average number of internal carbons in the dimer. The five outliers discussed in the previous section are denoted as diamonds. Color represents PAC type: green is an aliphatically linked PAC, red is an oxygenated PAC, and blue is a condensed hydrocarbon. Points with two colors share all the corresponding characteristics.

The plot of the number of internal carbons against the dimerization propensity, presented in Fig. 4, shows three somewhat distinct groupings: molecules with less than 10 internal carbons, which represent aliphatically linked hydrocarbons, pericondensed molecules with approximately 20 internal carbons, and larger pericondensed molecules with 30 or more carbons. When ignoring the outliers discussed in the previous section, these groupings generally correspond to the stability of the dimer, where aliphatically linked hydrocarbons are less stable than smaller pericondensed molecules and larger pericondensed molecules are the most stable, inline with previous works on the importance of shape of PACs [25] and on the lower dimerization speed and shorter lifetimes of linked PACs [21].

Finally, tessellation descriptors contain similar information of size and shape as they count the number of times four carbons are in proximity with each other (mostly internal carbons) and the number of times three carbons are in proximity with a hydrogen (mostly edge carbons).

3.2.2. Shape

The second group of properties corresponds to quantities that purely describe the shape of the molecules, such as the relative lengths of the first and second principal axis of inertia (WHIM mass axis 1 and 2 [32]), which are both size independent.

Figure 5 shows the ratio between the second and first principal axes of inertia (*i.e.*, aspect ratio), along with its relationship with the propensity of these molecules to dimerize. While a clear separation exists for the less stable dimers containing an aliphatically linked hydrocarbon, it is difficult to identify a trend for the remaining

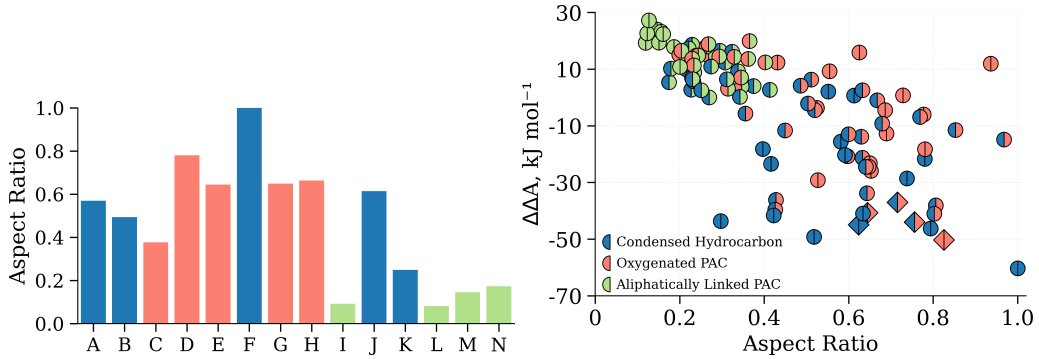

Fig. 5: Relationship between aspect ratio and dimerization FE. Left: The aspect ratio associated with each monomer. Right: Aggregation propensity compared to average aspect ratio in the dimer. The five outliers discussed in the previous section are denoted as diamonds. Color represents PAC type: green is an aliphatically linked PAC, red is an oxygenated PAC, and blue is a condensed hydrocarbon. Points with two colors share all the corresponding characteristics.

dataset. This suggests that size independent descriptors of shape are likely being used by the model only to identify aliphatically linked PACs and not other compounds. This phenomenon does not imply that shape descriptors do not have a clear relationship with the free energy, as they may play an important role for curved PACs [27]. However, due to the complexity of the FE landscape of curved molecules and the presence of multiple distinct configurations at short distances, we did not include any in this work.

3.2.3. Specific chemical groups

The third class of properties groups descriptors that are a metric for the presence of specific chemical groups, like the number of tessellations with three carbons and an oxygen atoms, the total Van der Waals surface area of all atoms with a partial charge between -0.05 and 0 (known as the vsa charge 7 [38]), and length of the longest aliphatic chain. The tessellation descriptor [33] considers each atom as a point in space and computes a Delaunay triangulation, counting the number of times each combination of four elements appear in a tessellation (see section 2 in the Supporting Information). While the property does not necessarily correspond directly to the number of oxygen atoms (an atom can appear in multiple tessellations), it accounts for the presence of oxygen by counting the number of times an oxygen is in proximity with three other carbons. The vsa charge 7 property encodes information about surface area but also implicitly captures information about oxygenated groups: most carbons that are located near oxygen functional groups are slightly positively charged and thus are not included in the surface area computation. Therefore, for equivalent sizes and geometries, the vsa charge 7 will be lower for molecules with electrophilic groups.

The last property in this group is the length of the longest aliphatic chain, which accounts for the presence of both rotatable bonds and side chains. In combination with the aspect ratio, this feature can distinguish between aliphatically linked chains

and side chains, which have the ability to stabilize PAC clusters and make aggregation more favorable [21, 20].

3.3. Hetero-aggregation

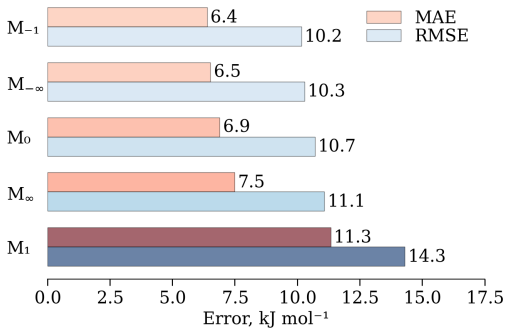

Fig. 6: Comparison of the predictive performance for different methods of combining monomer features for heterodimerization. $M_{-\infty}$ is the minimum value, M_{-1} is the harmonic mean, M_0 is the geometric mean, M_1 is arithmetic mean, and M_{∞} is the maximum value. For reference the RMSE of the input data (MD simulations) is $\sim 3.5\,\mathrm{kJ\,mol^{-1}}$.

Based on existing models, to predict the aggregation propensity for the heteromoelcular pairs, we computed the harmonic mean of the two monomers' molecular features. To test if this choice is optimal, we compared the performance of the model with five different combination rules. Using the definition of generalized mean,

$$M_p(x_1, \ldots, x_n) = \left(\frac{1}{n} \sum_{i=1}^{n} x_i^p \right)^{\frac{1}{p}} \tag{2}$$

we defined the combination rules as $M_{-\infty}$ (minimum value), M_{-1} (harmonic mean), M_0 (geometric mean), M_1 (arithmetic mean), and M_{∞} (maximum value).

The results, illustrated in Fig. 6, show that the harmonic mean outperforms the other metrics, even though the minimum value and geometric mean yield relatively similar results. This trend suggests that between two constituent molecules, the smaller properties tend to have a greater influence on the final stability. Interestingly however, the error (as a function of p) has a minimum, since $M_{-\infty} \leq M_{-1} \leq M_0$, but the difference is small enough for the current dataset that no further optimization is relevant. While in some cases (*e.g.*, charge or shape features) the magnitude of the property does not correspond to the size of the molecule, eight of the top ten features selected (see previous subsections) are extrinsic properties, suggesting that the characteristics of the smaller monomer plays a disproportionately larger role in the stability and the lifetime of the aggregate. This conclusion provides some empirical foundation to similar observations present in the literature [5, 26].

3.4. The effects of temperature

Up to this point, we considered only data at $1000\,\mathrm{K}$. While the current implementation of the model cannot be immediately extended to different temperatures, as much more data would be needed, we can test the generality of the selected features at different temperatures. Namely, we used the previously published homoaggregation FE obtained at $500\,\mathrm{K}$ and $1680\,\mathrm{K}$ [28] to train and test (at each temperature) a Lasso model using only the 10 features selected at $1000\,\mathrm{K}$. While the dataset covers a quite smaller subset of the data used at $1000\,\mathrm{K}$, at very different temperatures the balance of the entropic and enthalpic contributions differs, which can result in the aggregation giving more weight to different molecular characteristics. The prediction results at these two temperatures are shown in Fig. 7, with both temperatures, showing an RMSE and MAE lower than the one for the model trained on FE at $1000\,\mathrm{K}$, likely due to the smaller error associated with the prediction of homodimerization.

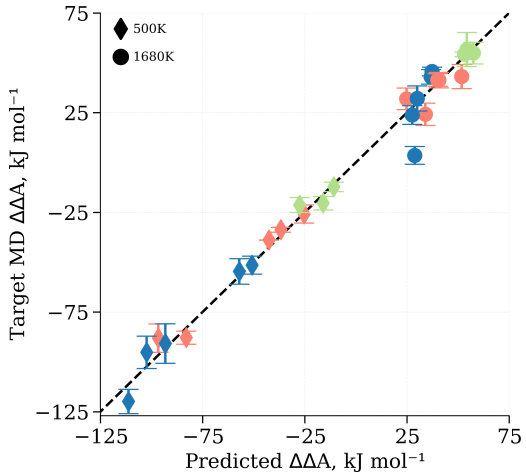

Fig. 7: Comparison of calculated (MD) and predicted FE of aggregation at $500\,\mathrm{K}$ (diamonds) and $1680\,\mathrm{K}$ (circles) using only the 10 features selected at $1000\,\mathrm{K}$. Color represents PAC type: green is an aliphatically linked PAC, red is an oxygenated PAC, and blue is a condensed hydrocarbon. Dashed line provides reference of correct predictions. At $500\,\mathrm{K}$, RMSE is $4.9\,\mathrm{kJ\,mol^{-1}}$ and MAE is $4.1\,\mathrm{kJ\,mol^{-1}}$. At $1680\,\mathrm{K}$ RMSE is $8.8\,\mathrm{kJ\,mol^{-1}}$ and MAE is $6.1\,\mathrm{kJ\,mol^{-1}}$.

Overall, the results show that the selected features are valid in a large temperature range. Of note, the error for the model trained with data at $1680\,\mathrm{K}$ is significantly greater than the one trained at $500\,\mathrm{K}$, potentially, because physical dimerization is a much less important process at this elevated temperature [11, 28] and the system tends towards the ideal gas behavior, for which many of the descriptors become meaningless.

4. Conclusions

In this work, we explored the relative importance of molecular properties on the physical aggregation of polycyclic aromatic compounds. As a first step, we created a dataset of the free energy of aggregation of PACs with different properties through enhanced sampling molecular dynamics. We then trained a Lasso based machine

learning model to predict the aggregation behavior of PACs, by representing each molecular pair with a set of molecular features. The model is able to find an accurate quantitative relationships between different properties of PACs and ultimately is able to predict the FE of dimerization with a higher degree of accuracy than traditional correlations.

The chosen method also provides insights into the properties important for dimerization by identifying the subset of features that are used to make the FE predictions. With no external bias towards specific properties, our model finds ten features that encode information that are related to PACs' size, shape, oxygenation, and presence of rotatable bonds. Critically, none of these properties is able, by itself, to capture the key aspects of PAC physical interactions, showing how the aggregation whether thermodynamic (pair stability) or kinetic (pair lifetime) is the result of the interplay of several properties. Finally, we show that the interactions between different PACs can be modeled by taking the harmonic mean of the individual species, although a larger variance in properties should be taken into account before generalizing to very large systems.

The findings of this paper offer insight into both PAC and soot formation as the stability effects of specific PAC features, not just their presence in the gas-phase, influence how commonly those features will be observed in the constituents of larger soot particles. Future works can shed further light on the nuances between these molecular properties and physical aggregation by expanding the dataset with new types of molecules such as curved PACs and developing novel features to numerically represent molecular properties in a chemically meaningful way.

Acknowledgments

The Authors thank Matt Raymond for the insightful discussions on machine learning methods. This work has been supported by the U.S Department of Transportation, FAA Center of Excellence 13-C-AJFE-GIT-067, the University of Michigan BlueSky Initiative, and the National Science Foundation Graduate Research Fellowship under Grant No. 1256260.

Supplementary material

Supplementary material associated with this article was submitted separately from this document.

References

[1] A. D'Anna, Combustion-formed nanoparticles, Proc. Combust. Inst. 32 (1) (2009) 593–613.
[2] H. Wang, Formation of nascent soot and other condensed-phase materials in flames, Proceedings of the Combustion Institute 33 (1) (2011) 41–67.
[3] M. R. Kholghy, G. A. Kelesidis, S. E. Pratsinis, Reactive polycyclic aromatic hydrocarbon dimerization drives soot nucleation, Phys. Chem. Chem. Phys. 20 (16) (2018) 10926–10938.
[4] J. H. Miller, The kinetics of polynuclear aromatic hydrocarbon agglomeration in flames, Symp. Combust. 23 (1) (1991) 91–98.

[5] J. D. Herdman, J. H. Miller, Intermolecular Potential Calculations for Polynuclear Aromatic Hydrocarbon Clusters, J. Phys. Chem. A 112 (28) (2008) 6249–6256.
[6] T. S. Totton, A. J. Misquitta, M. Kraft, A quantitative study of the clustering of polycyclic aromatic hydrocarbons at high temperatures, Phys. Chem. Chem. Phys. 14 (12) (2012) 4081–4094.
[7] A. Kazakov, H. Wang, M. Frenklach, Detailed modeling of soot formation in laminar premixed ethylene flames at a pressure of 10 bar, Combustion and Flame 100 (1-2) (1995) 111–120.
[8] P. Elvati, V. Dillstrom, A. Violi, Oxygen driven soot formation, Proc. Combust. Inst. 36 (1) (2017) 825–832.
[9] K. O. Johansson, T. Dillstrom, P. Elvati, M. Campbell, P. Schrader, D. Popolan-Vaida, N. Richards-Henderson, K. Wilson, A. Violi, H. Michelsen, Radical–radical reactions, pyrene nucleation, and incipient soot formation in combustion, Proc. Combust. Inst. 36 (1) (2017) 799–806.
[10] J. H. Miller, K. C. Smyth, W. G. Mallard, Calculations of the dimerization of aromatic hydrocarbons: Implications for soot formation, Symp. Combust. 20 (1) (1985) 1139–1147.
[11] Q. Mao, A. van Duin, K. Luo, Formation of incipient soot particles from polycyclic aromatic hydrocarbons: A ReaxFF molecular dynamics study, Carbon 121 (2017) 380–388.
[12] C. Schuetz, M. Frenklach, Nucleation of soot: Molecular dynamics simulations of pyrene dimerization, Proc. Combust. Inst. 29 (2) (2002) 2307–2314.
[13] S. Stein, A. Fahr, High-temperature stabilities of hydrocarbons, J. Phys. Chem. 89 (17) (1985) 3714–3725.
[14] J. Y. W. Lai, P. Elvati, A. Violi, Stochastic atomistic simulation of polycyclic aromatic hydrocarbon growth in combustion, Phys. Chem. Chem. Phys. 16 (17) (2014) 7969–7979.
[15] K. Johansson, J. Lai, S. Skeen, D. Popolan-Vaida, K. Wilson, N. Hansen, A. Violi, H. Michelsen, Soot precursor formation and limitations of the stabilomer grid, Proc. Combust. Inst. 35 (2) (2015) 1819–1826.
[16] J. Cain, A. Laskin, M. Kholghy, M. Thomson, H. Wang, Molecular characterization of organic content of soot along the centerline of a coflow diffusion flame, Phys. Chem. Chem. Phys. 16 (47) (2014) 25862–25875.
[17] M. Commodo, K. Kaiser, G. De Falco, P. Minutolo, F. Schulz, A. D'Anna, L. Gross, On the early stages of soot formation: Molecular structure elucidation by high-resolution atomic force microscopy, Combust. Flame 205 (2019) 154–164.
[18] J. Saldinger, Q. Wang, P. Elvati, A. Violi, Characterizing the diversity of aromatics in a coflow diffusion Jet A–1 surrogate flame, Submitted to Fuel (2019).
[19] J. C. Saldinger, P. Elvati, A. Violi, Stochastic and network analysis of polycyclic aromatic growth in a coflow diffusion flame, Phys. Chem. Chem. Phys. 23 (2021) 4326–4333.
[20] P. Elvati, A. Violi, Thermodynamics of poly-aromatic hydrocarbon clustering and the effects of substituted aliphatic chains, Proc. Combust. Inst. 34 (1) (2013) 1837–1843.
[21] S. Chung, A. Violi, Peri-condensed aromatics with aliphatic chains as key intermediates for the nucleation of aromatic hydrocarbons, Proc. Combust. Inst. 33 (1) (2011) 693–700.
[22] P. Elvati, A. Violi, Homo-dimerization of oxygenated polycyclic aromatic hydrocarbons under flame conditions, Fuel 222 (2018) 307–311.
[23] J. W. Martin, K. Bowal, A. Menon, R. I. Slavchov, J. Akroyd, S. Mosbach, M. Kraft, Polar curved polycyclic aromatic hydrocarbons in soot formation, Proc. Combust. Inst. 37 (1) (2019) 1117–1123.
[24] N. A. Eaves, S. B. Dworkin, M. J. Thomson, Assessing relative contributions of PAHs to soot mass by reversible heterogeneous nucleation and condensation, Proc. Combust. Inst. 36 (1) (2017) 935–945.
[25] A. Raj, M. Sander, V. Janardhanan, M. Kraft, A study on the coagulation of polycyclic aromatic hydrocarbon clusters to determine their collision efficiency, Combust. Flame 157 (3) (2010) 523–534.
[26] J. S. Lowe, J. Y. Lai, P. Elvati, A. Violi, Towards a predictive model for polycyclic aromatic hydrocarbon dimerization propensity, Proc. Combust. Inst. 35 (2) (2015) 1827–1832.
[27] Q. Wang, J. C. Saldinger, P. Elvati, A. Violi, Molecular structures in flames: A comparison between snaps2 and recent afm results, Proc. Combust. Inst. 38 (1) (2021) 1133–1141.
[28] P. Elvati, K. Turrentine, A. Violi, The role of molecular properties on the dimerization of aromatic compounds, Proc. Combust. Inst. 37 (1) (2019) 1099–1105.
[29] R. Tibshirani, Regression shrinkage and selection via the lasso, J. R. Stat. Soc. Ser. B Methodol. 58 (1) (1996) 267–288.
[30] J. C. Phillips, R. Braun, W. Wang, J. Gumbart, E. Tajkhorshid, E. Villa, C. Chipot, R. D. Skeel, L. Kalé, K. Schulten, Scalable molecular dynamics with NAMD, J. Comput. Chem. 26 (16) (2005) 1781–1802.
[31] G. A. Tribello, M. Bonomi, D. Branduardi, C. Camilloni, G. Bussi, PLUMED 2: New feathers for an old bird, Comput. Phys. Commun. 185 (2) (2014) 604–613.
[32] R. Todeschini, P. Gramatica, The Whim Theory: New 3D Molecular Descriptors for Qsar in Environmental Modelling, SAR and QSAR in Environmental Research 7 (1-4) (1997) 89–115.
[33] X. Yan, A. Sedykh, W. Wang, X. Zhao, B. Yan, H. Zhu, *In silico* profiling nanoparticles, Nanoscale 11 (17) (2019) 8352–8362.

[34] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, et al., Scikit-learn: Machine learning in python, Journal of machine learning research 12 (Oct) (2011) 2825–2830.
[35] L. M. Ghiringhelli, J. Vybiral, S. V. Levchenko, C. Draxl, M. Scheffler, Big Data of Materials Science: Critical Role of the Descriptor, Phys. Rev. Lett. 114 (10) (2015) 105503.
[36] M. Rapacioli, F. Calvo, F. Spiegelman, C. Joblin, Stacked Clusters of Polycyclic Aromatic Hydrocarbon Molecules, J. Phys. Chem. A 109 (11) (2005) 2487–2497.
[37] P. Elvati, E. Baumeister, A. Violi, Graphene quantum dots: effect of size, composition and curvature on their assembly, RSC Advances 29 (2017).
[38] P. Labute, A widely applicable set of descriptors, J. Mol. Graph 18 (4-5) (2000) 464–477.