Overriding Innate Decomposition Temperatures of an Avibactam Prodrug Precursor using Data Science-Guided Synthesis

Jacob Werth*, Michael Butler, Jenson Verghese, Nga Do, Lacey Samp, Remzi Duzguner and Michele Buetti-Weekly

Statistical analysis is used to correlate the thermal decomposition temperature of diverse leaving groups of an avibactam prodrug precursor. SMILES strings and Mordred calculated parameters were leveraged to provide a time-efficient workflow for model development. The resulting models were deployed to predict a novel analog with a higher onset temperature, allowing for an overall safer reagent and proof of concept for the workflow. Interperetation of the descriptors featured in the models and subsequent DFT analysis uncovered univariate trends providing a deeper understanding of the decomposition pathway. Finally, this workflow enabled the development of a predictive model correlating energy output of the precursor analogs for a more comprehensive assessment.

Introduction

Avibactam is an antibiotic used intravenously to treat extended spectrum beta-lactase infections.[1] Prodrugs of avibactam have been targeted as an orally delivered class to provide a more patient-centric therapy.[2] Through the development of this class, safety and logistical challenges of the prodrug precursor (**1**) in the final reaction with *N*-hydroxy (**2**) to the final API (**3**) was identified (Figure 1). Chlorosulfate **1** has a low Differential Scanning Calorimetry (DSC) onset (96 °C) and is an oil which requires low temperature storage (–20 °C) for quality assurance.

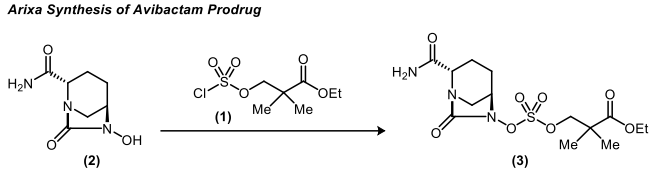

Fig 1. Overview of route to avibactam prodrug using a chlorosulfate precursor as previously disclosed.

DSC is an invaluable tool in assessing the initial safety profile of a novel compound.[3] In the context of process chemistry, it is imperative as a safeguard in de-risking scale-up experiments during the development process. Due to these undesirable attributes of **1**, an exploration to find alternative compounds that exhibit a higher DSC onset value, exist as a stable, crystalline solid and maintain reactivity towards **3** was undertaken. Initially, an empirical approach was used to screen various leaving groups in lieu of chloride including phenol, imidazole and pyrazole derivatives. Varied success was achieved through empirical efforts and the overall goal was not realized. This led us to apply data science tools to search for correlative models to extrapolate higher DSC values of a reactive class and further understand the effects of prodrug substitution. Although a large body of literature correlating DSC onset values of notoriously energetic functional groups (i.e. tetrazoles, azides, diazos) exists, there is limited precedent in using statistical analysis to predict DSC values of broad chemical functionality.[4-7] Herein, we present an ideal scenario, combining synthetic chemistry with computational tools, in which diverse screening data is leveraged as a training set to build multivariate linear regression (MLR) models capable of extrapolating DSC onset values. The computational workflow allows for rapid access to parameters and modelling. Further analysis of the key descriptors used for modelling provide mechanistic insight for the observed leaving group effects on thermal decomposition.

Results and Discussion

Synthesis of Analogs – Empirical Screening

Compound **1** was used to efficiently synthesize a diverse library of derivatives (Figure 2) through a variety of base-mediated conditions (see Supporting Information for details). The resulting compounds were tested using DSC to obtain onset temperatures. The values shown in Figure 2 represent the left-limit onsets which are the most appropriate for assessing the safety profile of new compounds.[8] We quickly found the DSC onset value of O-bound compounds to be broad and high (in some instances >200 °C). However, these were found to be unreactive under any conditions in productive synthesis of the desired avibactam prodrug **3**. N-bound heterocycles exhibited relatively lower DSC onset values but also did not produce desired product under basic conditions or with additives. Although the 2-Me-(benz)imidazole class of compounds were unreactive, they were of

Chemical Research & Development, Pfizer Worldwide Research and Development, Groton, Connecticut 06340, USA. Email: jacob.werth@pfizer.com
Footnotes relating to the title and/or authors should appear here.
Electronic Supplementary Information (ESI) available: [details of any supplementary information available should be included here].

https://doi.org/10.26434/chemrxiv-2023-9g0d5 **ORCID:** https://orcid.org/0000-0001-9648-5986 Content not peer-reviewed by ChemRxiv. **License:** CC BY-NC 4.0

particular interest as they were crystalline solids and exhibited relatively higher DSC onsets among N-bound heterocycles. Literature precedent of activating the 2-Me-imidazole and 2-Me-benzimidazole compounds with MeOTf was explored and the resulting compounds were found to be effective in producing avibactam prodrug in assay yields of 90-93% (Figure 3).[9, 10] The 2-Me-benimidazole analog was found to be the optimal compound considering the higher DSC onset value and comparable reactivity. Furthermore, the free base of this compound is stable at room temperature.

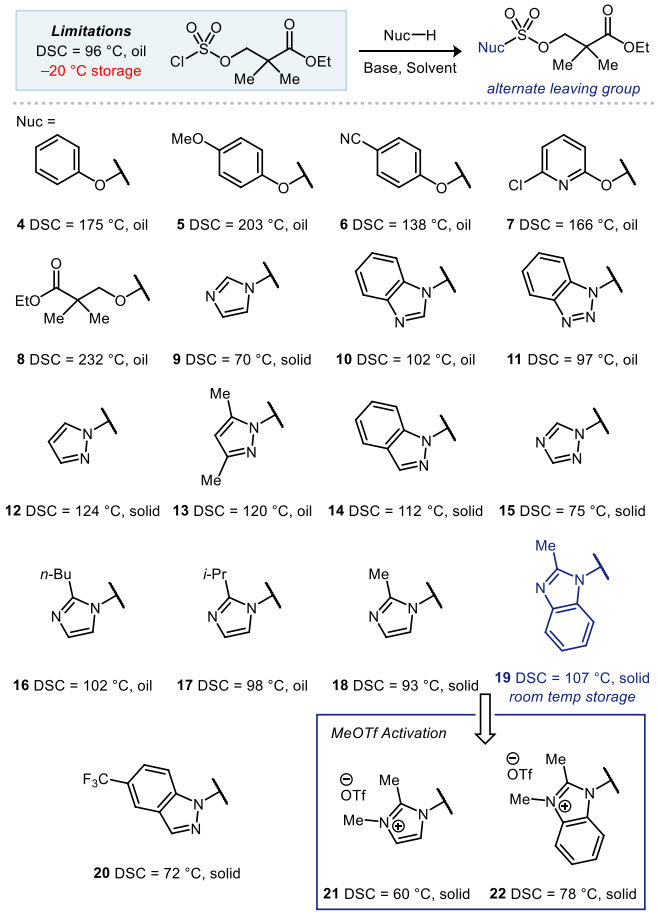

Fig 2. Overview of compounds synthesized to fit criteria for an alternate prodrug precursor leaving group.

Accelerated Rate Calorimetry (ARC) was performed on the 2-Me-benzimidazole analog and showed an onset at approximately 75 °C.[11] ARC is typically used as a follow-up experiment to measure the thermal stability of a compounds more accurately in comparison with DSC by using a slower, heat-wait-search ramp. The measured ARC onset temperature of the 2-Me-benzimidazole analog raised potential challenges in scale-up processes and future commercialization activities. At this point, we turned to data science tools to build correlative models in the hopes of understanding the relationship between leaving group and DSC onsets as well as predict a more thermally stable 2-Me-benzimidazole derivative.

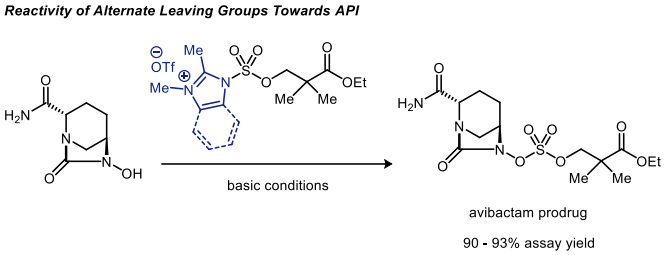

Fig 3. Reactivity towards avibactam prodrug using MeOTf salts of 2-Me-imidazole and 2-Me-benzimidazole leaving groups under basic conditions.

Machine Learning Approach

Our computational workflow began with screening our compounds through an internally built machine learning-based tool powered by ChemProp to assess initial prediction capability.[12] The model was trained on 1000's of Pfizer library compounds using only SMILES strings as input for parameter collection.[13] To our dismay, there was no correlation found between measured and predicted DSC onset values of the avibactam prodrug precursor analogs (Figure 4).

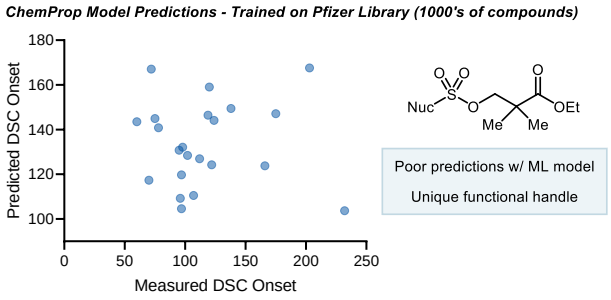

Fig 4. Measured vs. predicted DSC onsets of Chemprop model trained on Pfizer's library of compounds.

Considering the lack of representation for this specific compound class of substituted sulfates in the Pfizer library, we decided to alter our approach and develop de novo MLR models.[14] Based on the small training set of prodrug analogs, this analysis would take advantage of the chemical overlap of the prodrug between the compounds and rely on descriptors to capture the various leaving group effects.

MLR Model Development and Predictive Analysis

To begin our computational workflow, the entire set of analogs shown in Figure 2 (including **1**) were truncated by removing the prodrug component and taking forward the sulfonic acid derivatives as the training set (Figure 5). This was done to eliminate the conformational flexibility of the molecules for ease of analysis. Traditionally, MLR analysis of chemical structures often relies on conformational searching and density-functional theory (DFT)-level analysis to obtain sophisticated parameters relevant to the mechanism.[15] However, this approach requires a significant investment in time and computational resources.

https://doi.org/10.26434/chemrxiv-2023-9g0d5 ORCID: https://orcid.org/0000-0001-9648-5986 Content not peer-reviewed by ChemRxiv. License: CC BY-NC 4.0

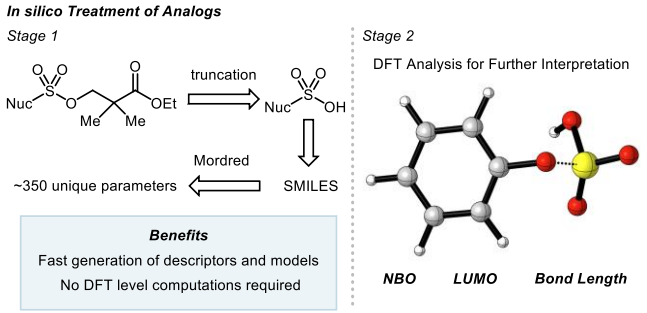

Fig 5. Overview of two-stage computational workflow used for MLR model development and mechanistic rationale.

In the interest of providing a user-friendly approach towards obtaining descriptors and expediting the timeline to initial modeling, SMILES strings were fed into the Mordred calculator using Python code.[16] Mordred is capable of producing 1000's of 2D and 3D parameters in seconds. After completion of Mordred calculations and cleaning of highly correlative parameters (R^2 > 0.95), approximately 350 features were taken forward for least squares regression analysis. This is in a similar vein with the machine learning algorithms mentioned above wherein DFT-level structures are not required for parameter collection. We envisioned DFT analysis could be deployed as a second stage of the workflow to help interpret unintuitive, quantitative structure activity relationship (QSAR) parameters and provide rationale for how a potential model is operative.

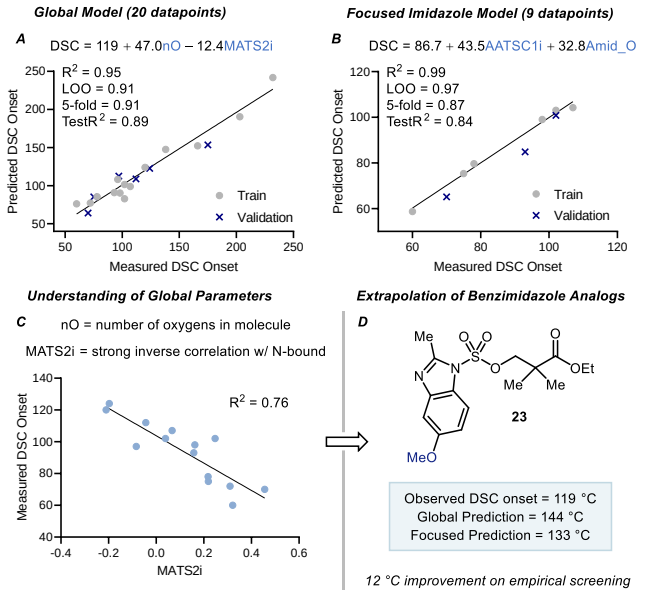

Fig 6. A. Global model for entire set of analogs. B. Focused model of imidazole-derived compounds. C. Preliminary understanding of parameters found in the Global model. D. Prediction of aryl-OMe substituted 2-Me-benzimidazole analog.

With the parameter library in hand, the data set was split into a training set/validation set of 70/30 and a forward stepwise algorithm was used to search for correlative models. Remarkably, several highly correlative (R^2 > 0.90), 2-term models were found. The optimal model for all the analogs is shown in Figure 6A. The additional statistical scores including LOO (leave-one-out) (0.91) and *k*-fold (*k* = 5, 0.91) indicated a robust model was found.[14] Furthermore, the test R^2 score of 0.89 provides a high level of confidence in predicting out of sample analogs. The most overwhelming type of parameter that showed up in the list of top models was a classification term representing the number of oxygen atoms present in the molecule. This represents a relatively simple classification but surprisingly provided an effective grouping of the structures. The second parameter used in the global model, moran coefficient of lag 2 weighted by ionization potential (MATS2i), which carries a negative coefficient is not as intuitive and by univariate analysis appears to have a more continuous role in correlating the compounds.

Further analysis of this parameter shows a highly correlative, inverse relationship with N-bound analogs and DSC onsets when O-bound analogs are omitted (Figure 6C). Qualitative analysis suggests the term is grouping the various N-bound heterocycle classes as well as capturing steric and electronic changes. Attempts to further understand this parameter was undertaken by DFT computation (M062X/def2TZVP//B3LYPD3BJ/6-31G(d,p)) of the truncated analogs and collection of more sophisticated, 3D descriptors. Parameters such as natural bond orbital (NBO) charges were collected from the lowest-energy conformer determined by DFT. No correlation (R^2 > 0.3) could be found between the MATS2i parameter and NBO charges, Sterimol-based parameters or other global descriptors (HOMO/LUMO energy, polarizability), highlighting the capacity for Mordred-based parameters to describe high-level features.[17]

In using the model to find more thermally stable compounds, the nO term was the focus for extrapolation. Given the simplicity of this type of parameter, this allows for a synthetic chemist to quickly search for compounds that would apply to this trend. Considering our interest in the reactive 2-Me-benzimidazole class, an aryl-OMe derivative was synthesized to test the model (Figure 6D). Along with an additional oxygen atom, this analog was computed and found to produce a MATS2i value of -0.30569, an extrapolation of the benzimidazole class. Gratifyingly, the compound exhibited a higher DSC onset value of 119 °C. The global model overpredicted the value by 25 °C.

Focused Model Development

The overprediction of the aryl-OMe derivative by the global model prompted us to explore a "focused" model comprised of (benz)imidazole-based compounds with the hypothesis that increasing the chemical overlap between structures may improve the prediction accuracy. Using 9 datapoints and a training set/validation set split of 60/40, the same MLR workflow used in the global model development was used to find an optimal 2-term model with overall high statistical scores (R^2 = 0.99, LOO = 0.97, 5-fold = 0.87) (Figure 6B). Using this focused model, the prediction error of the aryl-OMe analog was lowered to only 14 °C.

Interestingly, this model is comprised of 2 terms (Amid_O and AATSC1i) with positive coefficients. Univariate analysis of AATSC1i,

https://doi.org/10.26434/chemrxiv-2023-9g0d5 **ORCID:** https://orcid.org/0000-0001-9648-5986 Content not peer-reviewed by ChemRxiv. **License:** CC BY-NC 4.0

the parameter with the largest contribution, beside the measured DSC onsets show a strong correlation barring two clear outliers (Figure 7A). The two outliers were identified as the MeOTf salts. This trend indicates the second term in the model (Amid_O) is correcting for the MeOTf salt analogs. Further, when AATSC1i is compared with the collected DFT parameters, a strong correlation with Polarizability (R^2 = 0.92) is observed (Figure 7B).

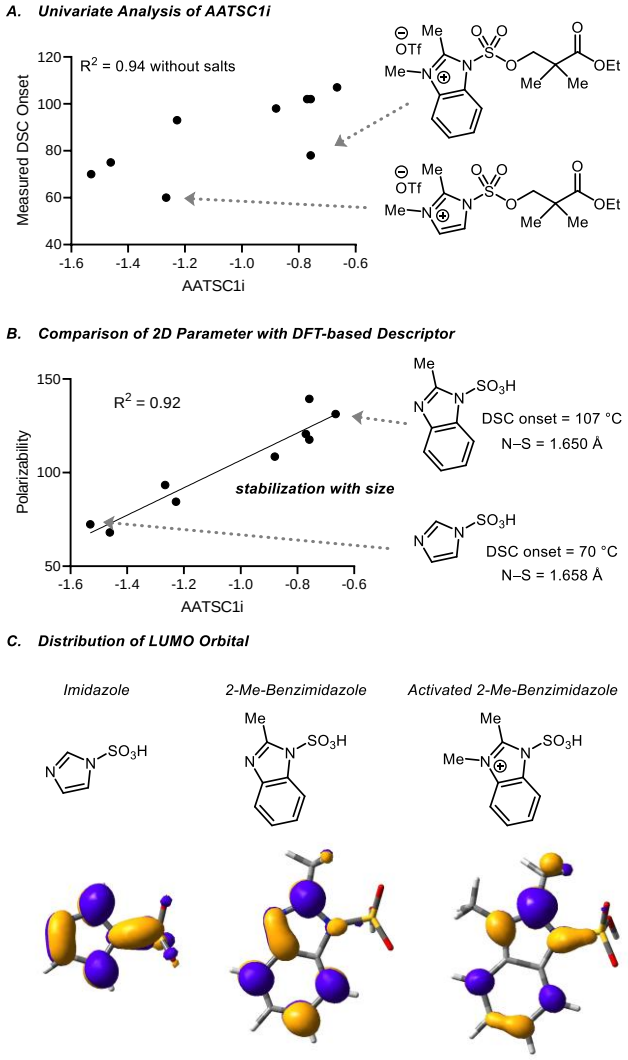

Fig 7. A. Univariate analysis of Measured DSC Onset and AATSC1i descriptor. B. Plot of AATSC1i (2D) and Polarizability (DFT-derived) parameters for imidazole-based compounds. C. Comparison of LUMO orbitals between imidazole, 2-Me-benzimidazole and MeOTf-activated 2-Me-benzimidazole truncated analogs.

Polarizability is a global parameter generally associated with overall size.[18] A high correlation between the AATSC1i parameter and molecular weight of the compounds was also discovered (R^2 = 0.89) Stabilization of the leaving group is seen with examples at each end of the spectrum for this parameter. Bond lengths of interest were calculated by optimizing the truncated analogs at the M062X/def2TZVP level of theory as this metric has been shown to be relevant in previous modelling studies of tetrazole DSC onsets.[6] Parent imidazole has a lower DSC onset at 70 °C and elongated N–S bond length of 1.658 Å while the larger 2-Me-benzimidazole analog has a significantly higher DSC onset at 107 °C and shorter N–S bond length of 1.650 Å. Qualitative comparison of the LUMO orbitals between the two analogs reveals a delocalization of the orbital from the N–S bond when extended conjugation is present which presumably is an additive effect with overall size stabilization. Additionally, these trends can rationalize the observed reactivity of activated 2-Me-benzimidazole towards API and lower DSC onset. The MeOTf salt of 2-Me-benzimidazole is measured computationally to have a N–S bond length of 1.712 Å and the LUMO orbital is redistributed across the N–S bond.

The improved predictive performance of the focused model speaks to the overall importance of chemical structure overlap in producing predictive models for unique functional groups. Ultimately, the 2-Me-OMe-benzimidazole analog would have unlikely been identified as a target without the knowledge gained from statistical analysis.

Effect of Aryl Electron Withdrawing Groups

We next tried to further increase the DSC onset value by introducing a methyl ester in lieu of the methoxy group (Figure 8A). This molecule was predicted to be much higher (Global = 192 °C, Focused = 164 °C) given the increase of oxygen atoms. Unfortunately, this compound exhibited a DSC onset of only 95 °C. Interestingly, the LUMO orbital of **24** is localized on the aryl carbonyl C–C bond as opposed to the S–O bond indicating an alternate decomposition pathway may have been introduced. This depression of onset value prompted us to take a closer look at the trends present in the compound library.

Interestingly, there appeared to be a negative aryl electron withdrawing group (EWG) effect present not only in the N-bound analogs but also the phenol-based O-bound ones. A stark example of this trend is present in a CF_3 substitution of an indazole analog resulting in a net loss of 40 °C from the parent indazole (Figure 8B). From DFT analysis, there was slight elongation of the N–S bond observed (1.662 Å for parent indazole vs. 1.668 Å for CF_3-Indazole) indicating another potential decomposition pathway may be operative. This aryl EWG trend is also evident among the phenol-based derivatives. A Hammett correlation between the para-substitution of the phenol ring and DSC onset values was observed and shows the detrimental effect of having an aryl EWG present on the ring (Figure 8C). This was an inverse trend relative to diazo compounds, which in a recent study were found to exhibit lower thermal stability with aryl electron-donating groups.[19] We hypothesize this could be the contributing reason for the aryl methyl ester compound showing a decrease in DSC onset value in comparison to the parent 2-Me-benzimidazole.

To further explore this Hammett trend, computational analysis of the S–O bond was carried out to understand potential elongation with increasing Hammett value/decreasing onset temperature. The 4-OMe analog S–O bond was computed and measured to be 1.578 Å. Comparatively, the 4-CN analog exhibited an S–O bond of 1.585 Å. This suggests the mechanism of decomposition is likely to occur at

https://doi.org/10.26434/chemrxiv-2023-9g0d5 ORCID: https://orcid.org/0000-0001-9648-5986 Content not peer-reviewed by ChemRxiv. License: CC BY-NC 4.0

the S–O bond of the leaving group. Visualization of the LUMO orbitals also displays higher relative density around the S–O bond in the 4-CN analog. Altogether, these trends identified with the aid of DFT analysis strengthen our understanding of the leaving group's effect on the prodrug precursor decomposition onset.

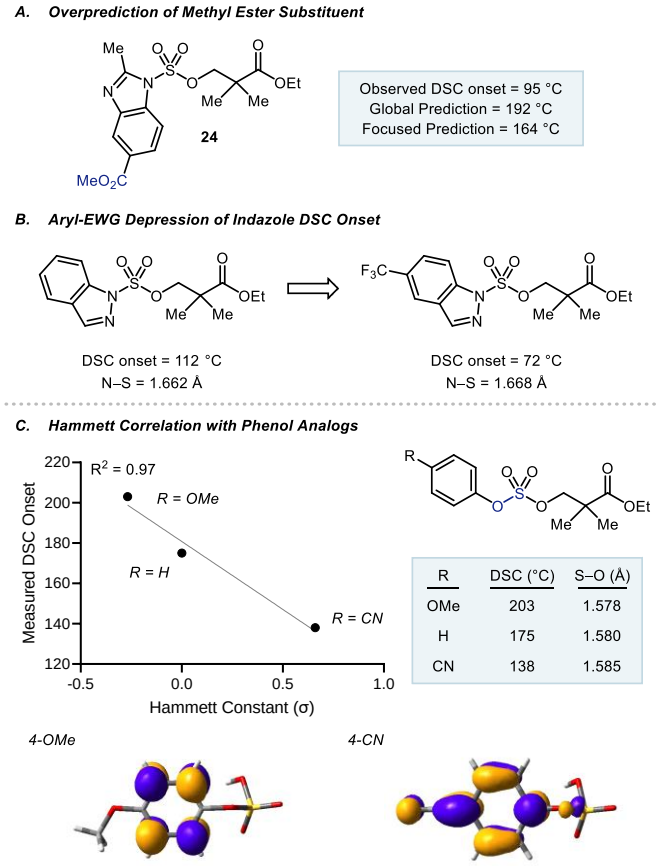

Fig 8. A. Prediction of aryl-CO_2Me substituted 2-Me-benzimidazole analog based on additional oxygen atoms. B. Identification of DSC depression of indazole-based analogs. C. Hammett correlation describing the inverse effect on phenol-based analogs. Analysis of computationally measured S–O bond length. Comparison of LUMO orbitals between 4-OMe and 4-CN phenol-based truncated analogs.

Model Development of Energy Output

Finally, we were interested in the potential to correlate our descriptor library with the energy output (J/g) of the DSC testing. The ability to predict the energy released from relative functional groups would provide a more comprehensive assessment of the thermal decomposition. Using the same workflow as introduced with DSC onset, initial attempts proved unfruitful in finding robust statistical models. Upon closer inspection, the chloride **(1)** and prodrug dimer **(8)** analogs were the main drivers for lack of correlation. Removal of these two compounds greatly improved the modeling efforts and resulted in multiple 3-term models with strong correlations. Given the necessity of 3 parameters and the removal of less commonly featured functional groups in the training set, this suggests energy is not as easily described as DSC onset temperature in the context of this study. The optimal model shown in Figure 9 scored high in statistical metrics (R^2 = 0.91, LOO = 0.83, 5-fold = 0.81, TestR^2 = 0.78).

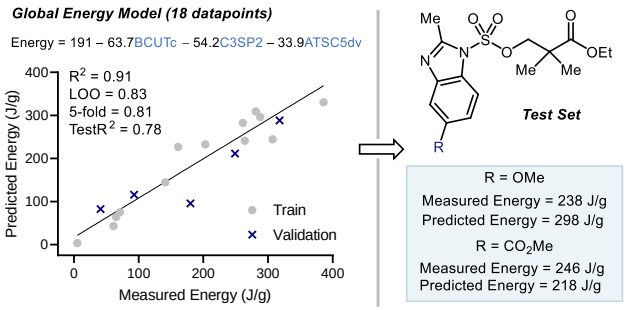

Fig 9. Global model correlating energy output from DSC experiments. External validation of the model using the 2-Me-benzimidazole derivatives as a test set.

Using the oxygenated, aryl 2-Me-benzimidazole derivatives as a test set, we sought to gauge the model's predictive capability. The model performed well in predicting the general energy output for the aryl-OMe (error: 60 J/g) and aryl-CO_2Me (error: 28 J/g) analogs. Considering the wide range of values in the study (0 – 400 J/g) and the mean average error (MAE) of the validation set (MAE = 43 J/g), we believe this to be a reasonable error for application.

Conclusion

In summary, a data science approach was used to correlate the thermal stability of disparate leaving groups for an avibactam prodrug precursor and further leveraged to predict an analog with a higher DSC onset value. By using a cost-effective, computational workflow, correlations were readily found and the time to prediction was significantly truncated. While 2D based-descriptors have traditionally been more difficult to interpret and attribute to mechanism due to their simplicity, this case study demonstrated how trends could be still be readily identified and overall provide a more holistic understanding of the prodrug precursor decomposition. Electronic contributions were discovered to influence the DSC onset values throughout the analogs and a Hammett correlation was defined among phenol-substituted derivatives which corresponded to lengthening of the S–O bond. Visualization of LUMO orbitals provided a qualitative reinforcement of the observed DSC trends. Overall, DFT analysis allowed for a more general understanding of a leaving group's effect on the prodrug precursor. As data science studies gravitate towards machine learning workflows and artificial intelligence with vast datasets, utilizing simple tools with smaller, curated datasets proves to still be a useful, complementary tool for chemists. We believe this workflow will provide a readily accessible entry point for nonexperts to apply data science tools in parallel with synthetic screening.

Conflicts of interest

There are no conflicts to declare.

https://doi.org/10.26434/chemrxiv-2023-9g0d5 **ORCID:** https://orcid.org/0000-0001-9648-5986 Content not peer-reviewed by ChemRxiv. **License:** CC BY-NC 4.0

Acknowledgements

The acknowledgements come at the end of an article after the conclusions and before the notes and references.

Notes and references

(1) Shirley, M. Ceftazidime-Avibactam: A Review in the Treatment of Serious Gram-Negative Bacterial Infections. *Drugs* **2018**, *78* (6), 675-692.
(2) Gordon, E. M.; Duncton, M. A. J.; Gallop, M. A. Orally Absorbed Derivatives of the β-Lactamase Inhibitor Avibactam. Design of Novel Prodrugs of Sulfate Containing Drugs. *J. Med. Chem.* **2018**, *61* (22), 10340-10344.
(3) Sheng, M.; Valco, D.; Tucker, C.; Cayo, E.; Lopez, T. Practical Use of Differential Scanning Calorimetry for Thermal Stability Hazard Evaluation. *Org. Process Res. Dev.* **2019**, *23* (10), 2200-2209.
(4) Beste, A.; Barnes, B. C. Prediction of thermal decomposition temperatures using statistical methods. *AIP Conf. Proc* **2020**, *2272* (1).
(5) Keshavarz, M. H.; Zohari, N.; Seyedsadjadi, S. A. Validation of improved simple method for prediction of activation energy of the thermal decomposition of energetic compounds. *J. Therm. Anal. Calorim.* **2013**, *114* (2), 497-510.
(6) Rein, J.; Meinhardt, J. M.; Hofstra Wahlman, J. L.; Sigman, M. S.; Lin, S. A Physical Organic Approach towards Statistical Modeling of Tetrazole and Azide Decomposition**. *Angew. Chem. Int. Ed.* **2023**, *62* (17), e202218213.
(7) Zohari, N.; Abrishami, F.; Zeynali, V. Prediction of decomposition temperature of azole-based energetic compounds in order to assess of their thermal stability. *J. Therm. Anal. Calorim.* **2020**, *141* (4), 1453-1463.
(8) Green, S. P.; Wheelhouse, K. M.; Payne, A. D.; Hallett, J. P.; Miller, P. W.; Bull, J. A. On the Use of Differential Scanning Calorimetry for Thermal Hazard Assessment of New Chemistry: Avoiding Explosive Mistakes. *Angew. Chem. Int. Ed.* **2020**, *59* (37), 15798-15802.
(9) Beaudoin, S.; Kinsey, K. E.; Burns, J. F. Preparation of Unsymmetrical Sulfonylureas from N,N'-Sulfuryldiimidazoles. *J. Org. Chem.* **2003**, *68* (1), 115-119.
(10) Lee, H. K.; Bang, M.; Pak, C. S. Efficient synthesis of arylsulfamides by reaction of amines with arylsulfamoyl imidazolium triflate. *Tetrahedron Lett.* **2005**, *46* (42), 7139-7142.
(11) Sheng, M.; Valco, D.; Tucker, C. Heat Loss in Accelerating Rate Calorimetry Analysis and Thermal Lag for High Self-Heat Rates. *Org. Process Res. Dev.* **2021**, *25* (1), 108-119.
(12) Yang, K.; Swanson, K.; Jin, W.; Coley, C.; Eiden, P.; Gao, H.; Guzman-Perez, A.; Hopper, T.; Kelley, B.; Mathea, M.; et al. Analyzing Learned Molecular Representations for Property Prediction. *J. Chem. Inf. Model.* **2019**, *59* (8), 3370-3388.
(13) Weininger, D. SMILES, a chemical language and information system. 1. Introduction to methodology and encoding rules. *J. Chem. Inf. Comput. Sci.* **1988**, *28* (1), 31-36.
(14) Santiago, C. B.; Guo, J.-Y.; Sigman, M. S. Predictive and mechanistic multivariate linear regression models for reaction development. *Chem. Sci.* **2018**, *9* (9), 2398-2412.
(15) Crawford, J. M.; Kingston, C.; Toste, F. D.; Sigman, M. S. Data Science Meets Physical Organic Chemistry. *Acc. Chem. Res.* **2021**, *54* (16), 3136-3148.
(16) Moriwaki, H.; Tian, Y.-S.; Kawashita, N.; Takagi, T. Mordred: a molecular descriptor calculator. *J. Cheminf.* **2018**, *10* (1), 4.
(17) Brethomé, A. V.; Fletcher, S. P.; Paton, R. S. Conformational Effects on Physical-Organic Descriptors: The Case of Sterimol Steric Parameters. *ACS Catal.* **2019**, *9* (3), 2313-2323.
(18) Werth, J.; Sigman, M. S. Connecting and Analyzing Enantioselective Bifunctional Hydrogen Bond Donor Catalysis Using Data Science Tools. *J. Am. Chem. Soc.* **2020**, *142* (38), 16382-16391.
(19) Green, S. P.; Wheelhouse, K. M.; Payne, A. D.; Hallett, J. P.; Miller, P. W.; Bull, J. A. Thermal Stability and Explosive Hazard Assessment of Diazo Compounds and Diazo Transfer Reagents. *Org. Process Res. Dev.* **2020**, *24* (1), 67-84.

https://doi.org/10.26434/chemrxiv-2023-9g0d5 **ORCID:** https://orcid.org/0000-0001-9648-5986 Content not peer-reviewed by ChemRxiv. **License:** CC BY-NC 4.0