A Neural Network Model Informs Total Synthesis of Clovane Sesquiterpenoids

Pengpeng Zhang[†,1], Jungmin Eun[†,1], Masha Elkin[1], Yizhou Zhao[1], Rachel L. Cantrell[1], and Timothy R. Newhouse[1,*]

[1]Department of Chemistry, Yale University, 225 Prospect St., New Haven, CT, 06511, USA.

[†]These authors contributed equally

*Corresponding author. Email: timothy.newhouse@yale.edu

Abstract: Efficient syntheses of complex small molecules often involve speculative experimental approaches. The central challenge of such plans is that experimental evaluation of high-risk strategies is resource intensive, as it entails iterative attempts at unsuccessful strategies. Herein, we report a complementary strategy that combines creative human-generated synthetic plans with robust computational prediction of the feasibility of key steps in the proposed synthesis. A neural network model was developed to predict the outcome of a generally disfavored transformation, the 6-*endo*-*trig* radical cyclization, and applied to synthetic planning of clovan-2,9-dione, resulting in a 5-step total synthesis that improves on a prior 15-step approach. This work establishes how machine learning can guide multistep syntheses that employ innovative and high-risk human-generated plans.

One-Sentence Summary: A machine learning model was developed to predict the yield of a chemical transformation and guide the total synthesis of complex small molecules.

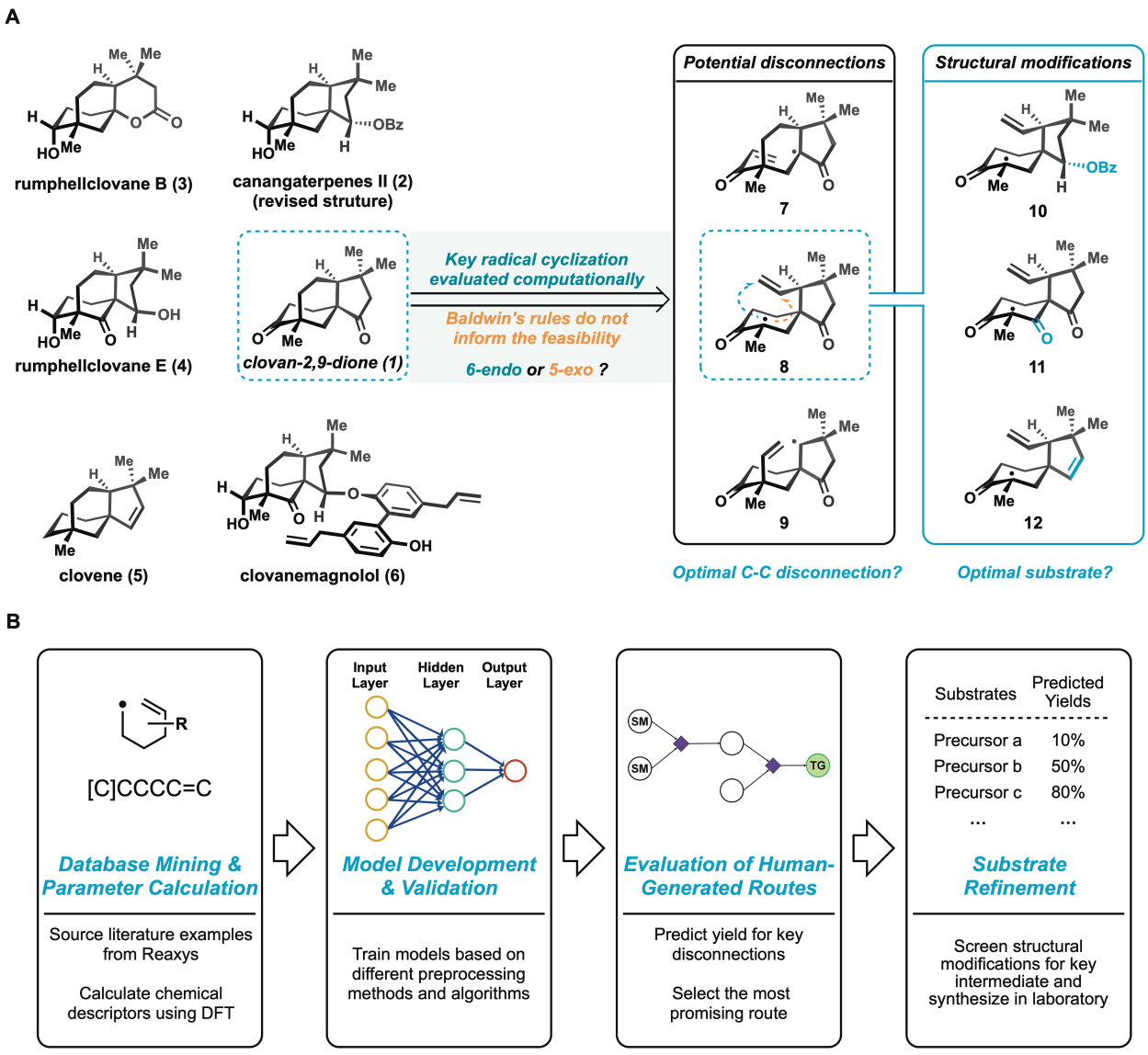


Fig. 1. ML model informs synthetic plan for clovane sesquiterpenoids. (**A**) Clovane-type natural product synthesis proposed via 6-*endo-trig* radical cyclizations, but the feasibility and optimal substrate were uncertain. (**B**) Workflow for the development and application of a machine learning model to guide synthetic planning.

The synthesis of small molecules is integral to a variety of disciplines, from materials science to medicinal chemistry. For complex small molecules, efficient chemical synthesis requires detailed retrosynthetic planning (*1*) and experimental evaluation. These plans usually involve one or more key steps that generate significant structural complexity. When key steps initially fail, different iterations of the key step are attempted, which is time and resource intensive to the extent that strategies are sometimes abandoned. This process has unfortunately been necessary as nuanced changes in substrate structure often result in significant changes in chemical reactivity that are challenging to predict.

One exciting approach to address the challenges associated with synthetic design is computer-aided retrosynthetic planning (*2-5*) wherein computational approaches are used to

provide synthetic routes. Herein, we report a complementary strategy that combines creative human-generated synthetic plans with robust computational analysis to predict the feasibility of key steps in proposed syntheses. Specifically, we report the development of a neural network model that is used to evaluate human-generated synthetic strategies towards clovane sesquiterpenoids by predicting the yields of key 6-*endo*-*trig* radical cyclization steps (Figure 1A). Our approach uniquely integrates computation into human retrosynthetic analysis through iterative virtual screening.

Radical cyclizations constitute a powerful method for the construction of hindered or strained ring systems and are commonly employed in complex molecule synthesis (*6*). However, the utility of some radical cyclizations, including often unfavorable *6-endo* cyclizations, can be limited by the difficulty of predicting their outcome. Baldwin's and Beckwith's rules (*7*) and other methods of analysis can in some cases suggest trends for related systems, but cannot quantitatively inform the outcome of diverse proposed transformations. A more sophisticated prediction of synthetic feasibility is enabled by the machine learning (ML) model described herein, which was applied to the synthetic planning of clovane sesquiterpenoids as a proof-of-concept of this approach. The clovanes share a common tricyclic bridged-ring skeleton with three quaternary centers, and have been a subject of synthesis since the 1960s (*8*, *9*). Additionally, clovanes exist widely in both terrestrial and marine organisms (*10*) and show diverse biological activities: for example, clovan-2,9-dione (**1**) and rumphellclovane B (**3**) inhibit production of superoxide anion and inhibit elastase released by human neutrophils (*11, 12*); clovanemagnolol (**6**) exhibits excellent neurotrophic activity at concentrations of 10 nM (*13*). Stunning biomimetic semi-syntheses have been reported of clovanes (*13, 14*), but semi-synthetic approaches provide limited opportunity for deep-seated structural modifications and concise access to related materials (*15*). Our de novo synthesis reported herein provides flexible entry and access to diverse clovanes that complement those available from semi-synthetic approaches.

To develop and apply a ML model to complex molecule synthesis, we devised the following workflow (Figure 1B): (1) a library of literature examples was collected and annotated with chemical descriptors from simple and readily conducted DFT calculations (*16*); (2) different machine learning model architectures were trained and evaluated for predictive performance; (3) human-generated retrosynthetic disconnections were evaluated using the trained ML model; (4) for the selected disconnection, substituents and functional groups were virtually screened with the model.

The feasibility of using machine learning to enable the total synthesis of clovanes is supported by complementary research in synthetic methods development using chemoinformatics (*17-20*). These workflows inspired our efforts, but none of them could be directly applied to complex molecule synthesis. The major differences are summarized here: (a) the substrates used in synthetic methodology development are readily available, whereas substrates involved in complex molecule synthesis require time consuming multistep synthetic operations to obtain; (b) similar substrates, ligands, or catalysts often appear in multiple instances throughout the libraries used for synthetic methodology, which cover a relatively narrow region of chemical space, whereas the substrates and products in our radical cyclization library are highly diverse; (c) the datasets generated from a single source (i.e. high-throughput experimentation) or a small number of literature references are relatively homogenous, whereas our datasets are derived from highly heterogenous sources with considerable variability in protocol and reaction conditions.

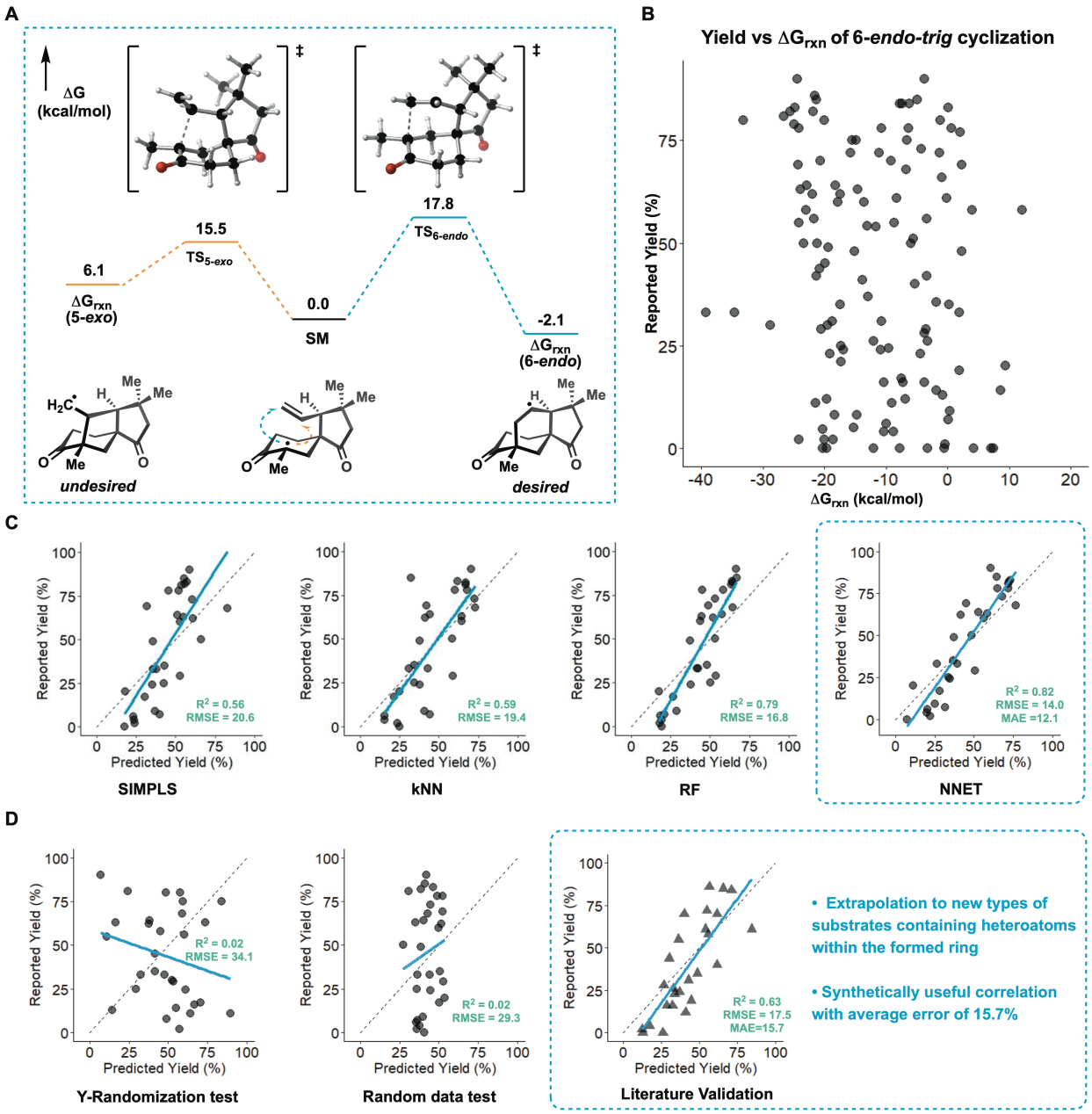


Fig. 2. ML model development. (**A**) DFT calculations for radical cyclization towards clovan-2,9-dione (uB3LYP/6-311++(d,p)). (**B**) Computed free energies of 6-*endo-trig* radical cyclization (ΔG_{rxn}) do not correlate with cyclization yields. (**C**) Performance of different ML models on test data set. (**D**) Control experiments and literature validation for the optimal NNET model.

Although a purely DFT approach was successful for substrate selection in the case of the total synthesis of paspaline A and emindole PB (*21*), methods that evaluate energies of products and transition states would be challenging for this radical cyclization due to a complex interrelationship between kinetics and thermodynamics. As can be seen in Figure 2A, the *5-exo* mode of cyclization is kinetically favored whereas the desired *6-endo* radical cyclization intermediate is thermodynamically favored; it was unknown if greater thermodynamic preference (ΔG_{rxn}) would result in higher yield of the *6-endo* product (*22*). To investigate this possibility, the experimental yields of 125 literature reactions were plotted against their computed free energies

of reaction (ΔG_{rxn}). The lack of a correlation suggests that yield is determined by many factors besides ΔG_{rxn}. It was thus hypothesized that a multiparameter ML model would allow for accurate yield predictions of 6-*endo-trig* radical cyclizations, which was needed to evaluate synthetic feasibility.

With this hypothesis in mind, we first obtained a library of literature examples of 6-*endo-trig* radical cyclizations from Reaxys®. Reactions were limited to sp^3-centered radicals undergoing intramolecular cyclization onto a pendant olefin, resulting in a set of 99 reactions, which include a fairly even distribution of yields from 0 to 90%. For each reaction in the library, radical intermediates before and after cyclization were subjected to simple and rapid DFT calculations (uB3LYP/6-31g(d)) of physical descriptors (*16*). A total of 340 descriptors per reaction were extracted to constitute the input parameters, including molecular, atomic, steric descriptors and linear combinations (see SI for details). Next, the library was split into training and test datasets (70/30) by the Kennard-Stone sampling method (*20*). As a large number of descriptors (340) were used relative to the small library size (99), overfitting was a significant concern. Therefore, feature selection with correlation filtering and PCA dimension reduction (*23*) were employed to transform 340 descriptors into 20 orthogonal parameters.

An array of supervised ML models was tuned with 10-fold cross-validation on training data and then were evaluated against the test dataset to provide R^2 and RMSE (root mean square error) values. As shown in Figure 2C, SIMPLS (Statistically Inspired Modification of the Partial Least Squares), and kNN (k-Nearest Neighbors) algorithms showed moderate predictive performance on the test dataset with R^2 values of 0.56 and 0.59, respectively. A random forest (RF) model provided better performance with $R^2 = 0.79$. A single hidden layer neural network (NNET) delivered improvement over these methods, providing a R^2 value of 0.82, with RMSE of 14.0% and MAE (mean absolute error) of 12.1%.

To evaluate the soundness of our NNET model, two control experiments were conducted: Y-randomization, in which yields are randomly shuffled across the dataset; and a random data test, where chemically meaningful descriptors are replaced with randomly generated values (*24*). The low correlations observed ($R^2 = 0.02$) suggest that the predictions of our NNET model were achieved by identifying relationships between yield and chemically meaningful featurization, rather than by finding chance correlations. To test the model's ability to extrapolate beyond the template library, literature validation was conducted with an additional 26 examples of 6-*endo* radical cyclization from Reaxys® and SciFinder®; these substrates contained functional groups that are not represented in the training or testing datasets, such as heteroatoms (N, O) within the formed 6-membered ring (see SI for details). We were pleased to find that reasonable correlation was observed, even though the model was not trained on these types of substrates. The lower correlation ($R^2 = 0.63$) and higher MAE (15.7%) are likely due to the different chemical reactivity between substrates possessing all-carbon skeletons and those with heteroatoms. For the purposes of clovane sesquiterpenoid synthesis, it was not necessary to have high performance for these substrate types, but the reasonable performance suggests our model provides synthetically useful predictions.

With the trained NNET model, different disconnections corresponding to different synthetic routes to clovan-2,9-dione (**1**) were evaluated (Figure 3A). The predicted yields of 6-*endo-trig* radical cyclizations from precursors **7**, **8**, and **9** are 46%, 46%, and 34%, respectively. The synthetically useful yield predicted for compound **7** is expected as such a disconnection with a polarized alkene is the conventional strategy for eliciting the 6-*endo* mode of cyclization. Although conventional logic would have discouraged the selection of **8** and **9** due to limited available

precedent for cyclizations of this type (*25*), the model's encouraging predictions for **8** and **9** mitigated that concern and neutralized any perceived benefit to the lower risk strategy via **7**. Finally, compound **8** was selected, as **8** has a comparable predicted yield and represents a more innovative disconnection (*9*) that leads to greater synthetic accessibility (*26*).

The next consideration regarded which proximal and remote functionality would be the optimal choice for the substrate given synthetic accessibility, predicted efficiency, and utility in accessing a variety of clovanes. A selection of substrates from over 100 predictions is shown in Figure 3A to illustrate the planning considerations that were made. For example, the introduction of carbonyl groups stabilizing the radical intermediate in triketone **13**, the protection of carbonyl group adjacent to the radical center in **14**, and the introduction of substituents at the site of radical cyclization in **15**, have predicted yields that are qualitatively in line with expert intuition. Compounds **16** and **17**, which would readily lead to other clovane natural products, are predicted to cyclize in synthetically useful yields.

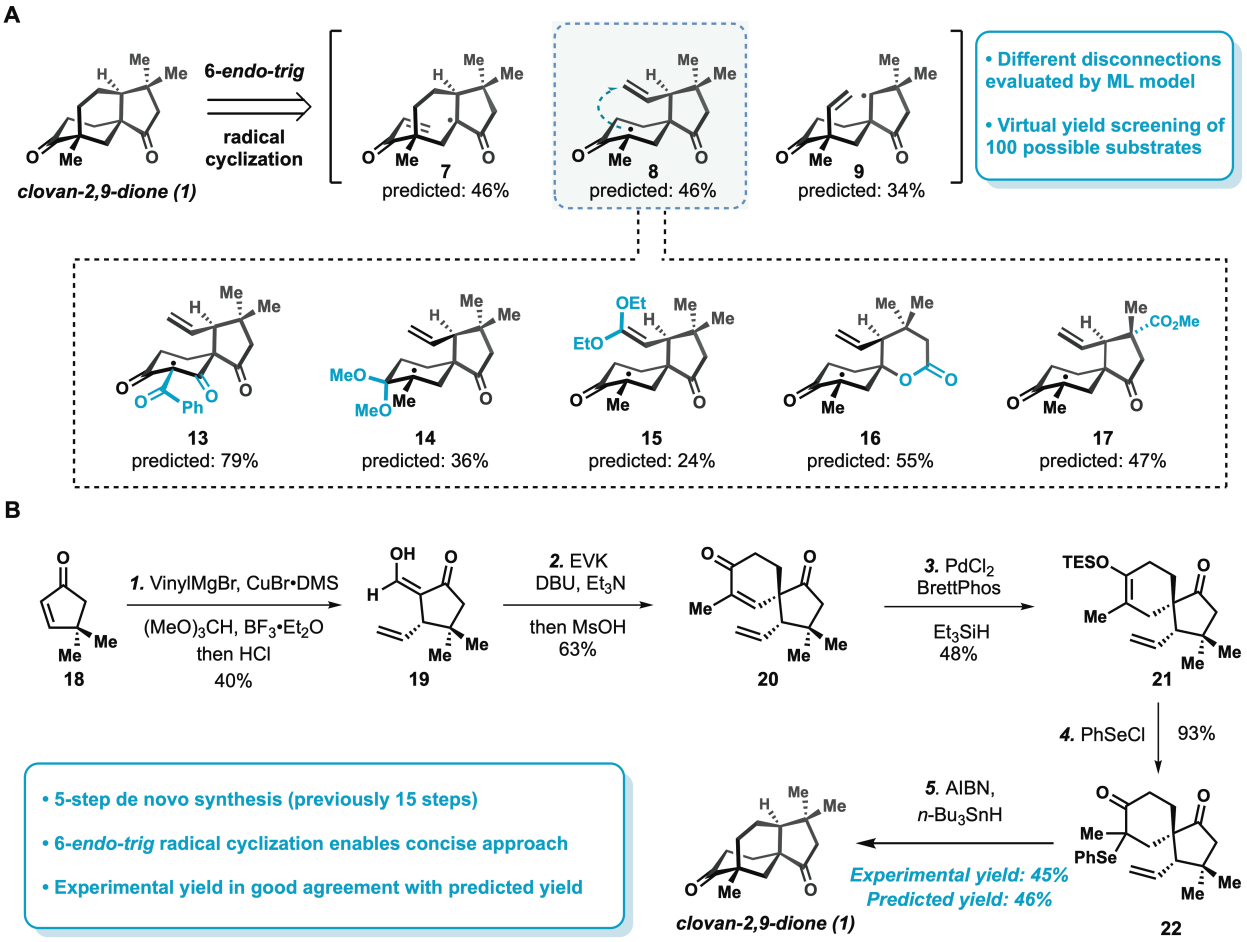
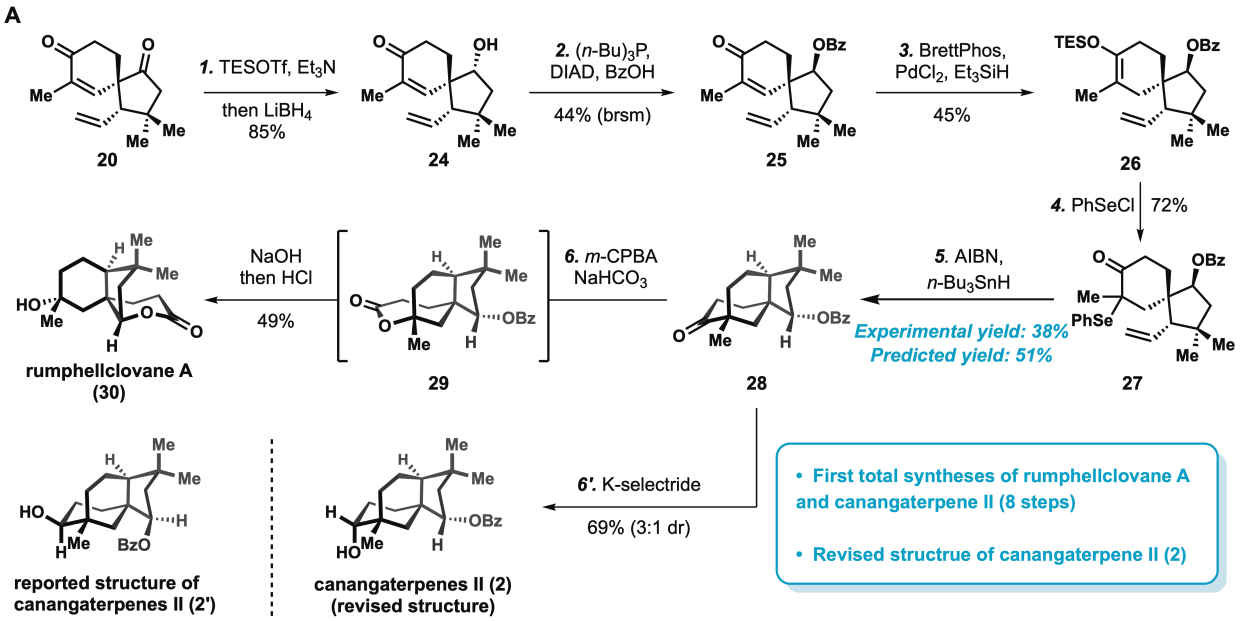


Fig. 3. ML model informed synthesis of clovan-2,9-dione. (**A**) ML-guided retrosynthetic analysis and substrate refinement. (**B**) Counterintuitive 6-*endo-trig* radical cyclization enables the 5-step total synthesis of clovan-2,9-dione (**1**).

As shown in Figure 3B, the synthetic route via radical intermediate **8** to clovan-2,9-dione (**1**) starts from commercially available 4,4-dimethylcyclopent-2-en-1-one (**18**). Vicinal difunctionalization with a vinyl cuprate nucleophile and enolate trapping with $HC(OMe)_3$ provided **19**. Adduct **19** underwent a Robinson annulation with ethyl vinyl ketone (EVK) to afford **20**. An

enone-selective Pd-catalyzed hydrosilylation of **20** (*27*) provided **21**. These newly developed conditions were necessary as Pt, Cu, and Rh catalysts provided inferior results. Treatment of the enoxysilane **21** with PhSeCl provided the radical precursor **22** as an inconsequential mixture of diastereomers.

When **22** was subjected to optimized radical cyclization conditions (AIBN, *n*-Bu_3SnH), clovan-2,9-dione (**1**) was produced in 45% yield. This result is in excellent agreement with the predicted yield of 46%. Since the literature examples used to train the NNET model are generally optimized yields, the predictions are for optimized yields as well. The successful realization of this radical cyclization resulted in a 5-step synthesis of **1**, which is significantly more efficient than the previously disclosed 15-step strategy (*8*).



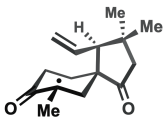
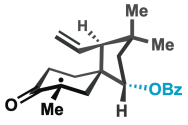
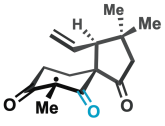
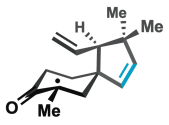
Substrate	 8	 10	 11	 12
ML predicted yield	46%	51%	62%	42%
Experimental yield	45%	38%	68%	49%

Fig. 4. Experimental validation of NNET model and application to the total synthesis of clovane sesquiterpenoids. (**A**) The first total syntheses of rumphellclovane A (**30**) and canangaterpene II (**2**). (**B**) Experimental validation of the NNET model.

As shown in Figure 4A, the NNET model predicted the feasibility of radical cyclization from **27** to form **28** (predicted yield: 51%). The experimental success of this transformation enabled the first total syntheses of rumphellclovane A (**30**) (*28*) and canangaterpene II (**2**) (*29*) in 8 steps from commercially available **18**. The key elements of the synthesis are selective reduction of **20** to form **24**, a Mitsunobu reaction to invert the stereochemistry to **25**, and a late-stage Baeyer-Villiger and

selective transesterification (**28** to **30**). The structure of canangaterpene II (**2**) was revised from the previously proposed structure based on biosynthetic considerations (*14*), NMR calculations, and our synthesis of the revised structure (see SI for details).

To rigorously test the model performance, we examined four radical precursors (**8, 10**–**12**) as an experimental validation set. As shown in Figure 4B, the experimental yields are in excellent agreement with the predicted yields, demonstrating the validity of our model. Accurate DFT calculations of the full pathway for >100 substrates would be computationally intractable for the time scales necessary for retrosynthetic analysis. With the model reported herein, dozens of substrates can be evaluated in less than a day, whereas full analysis by DFT of a single substrate requires weeks.

In summary, this report describes a platform that combines creative human-generated synthetic plans with robust computational analysis for a challenging key step. Machine learning models are trained to predict the yields of reactions from diverse literature examples. A neural network model was used to guide the retrosynthetic analysis of several sesquiterpenoid natural products, resulting in their highly efficient syntheses. We expect that models for other transformations could be developed following this workflow, which would allow for evaluation of retrosynthetic plans with varying key transformations. Moreover, the success of this strategy argues for broader use of computational tools as part of the process for synthetic planning.

References and Notes

1. E. J. Corey, *Chem. Soc. Rev*. **17**, 111–133 (1988).
2. Y. Shen *et al*., *Nat. Rev. Methods Primers* **1**, 23 (2021).
3. C. W. Coley *et al.*, *Angew. Chem. Int. Ed.* **59**, 22858–22893 (2020).
4. M. H. S. Segler *et al*., *Nature* **555**, 604–610 (2018).
5. S. Szymkuc *et al*., *Angew. Chem. Int. Ed.* **55**, 5904–5937 (2016).
6. K. J. Romero *et al*., *Chem. Soc. Rev.* **47**, 7851–7866 (2018).
7. J. E. Baldwin, *J. Chem. Soc., Chem. Commun.* **18**, 734–736 (1976).
8. J. Yang *et al*., *Org. Lett*. **19**, 6040−6043 (2017).
9. G. Liu *et al*., *Org. Lett*. ***23***, 290–295 (2021).
10. F. L. Bideau *et al*., *Chem. Rev*. **117**, 6110−6159 (2017).
11. H.-M. Chung *et al*., *Bull. Chem. Soc. Jpn*. **84**, 119–121. (2011).
12. H.-M. Chung *et al*., *Tetrahedron Lett*. **69**, 2740–2744 (2013).
13. X. Cheng *et al*., *Org. Biomol. Chem.* **10**, 383–393 (2012).
14. G. G. de Souza *et al*., *Org. Biomol. Chem*. **10**, 3315–3320 (2012).
15. P. M. Wright *et al*., *Angew. Chem. Int. Ed*. **53**, 8840–8869 (2014).
16. K. Jorner *et al*., *Nat. Rev. Chem.* **5**, 240–255 (2021).
17. J. P. Reid, M. S. Sigman, *Nature* **571**, 343–348 (2019).
18. D. T. Ahneman *et al*., *Science* **360**, 186–190 (2018).

19. B. J. Shields *et al.*, *Nature* **590**, 89–96 (2021).
20. A. F. Zahrt *et al*., *Science* **363**, eaau5631(2019).
21. D. E. Kim *et al*., *J. Am. Chem. Soc*. **141**, 1479–1483 (2019).
22. A. L. J. Beckwith, C. H. Schiesser, *Tetrahedron* **41**, 3925–3941 (1985).
23. T. A. Wenderski *et al*., *Methods Mol. Biol.* **1263**, 225–242 (2015).
24. J. G. Estrada *et al*., *Science* **362**, eaat8763 (2018).
25. V. Valerio *et al*., *Chem. Eur. J.* **22**, 2640–2647 (2016).
26. J. Siewert *et al*., *Chem. Eur. J*. **13**, 7424–7431 (2007).
27. M. Benohoud *et al*., *Chem. Eur. J*. **17**, 8404–8413 (2011).
28. H.-M. Chung, *et al*., *Tetrahedron Lett.* **51**, 2734–2736 (2010).
29. T. Matsumoto, *et al*., *J. Nat. Prod*. **77**, 990–999 (2014).

Acknowledgments:

We gratefully acknowledge Yale University's High-Performance Computing (HPC) Center for providing resources for this work. Dr. Fabian Menges is gratefully acknowledged for obtaining the high-resolution mass spectrometry data.

Funding:

We are grateful for financial support from Yale University, the Sloan Foundation, Boehringer Ingelheim, Genentech, and the NIH (GR100045). Student support included Chemical Biology Training Grant (T32 GM067543 to RLC) and an Anderson Postdoctoral Fellowship (PZ).

Author contributions:

ME, PZ and TRN initiated the project. JE, PZ, TRN, YZ designed the synthetic routes; JE, PZ, RLC, YZ carried out the chemical synthesis, JE carried out the DFT calculations; ME, PZ, and YZ carried out the ML modeling; all co-authors wrote and edited the manuscript.

Competing interests: Authors declare that they have no competing interests.

Data and materials availability: All data are available in the main text or the supplementary materials.

Supplementary Materials

Experimental Details of Total Syntheses

Supplementary Text

Fig. S1

Tables S1 to S3

NMR Spectra

The Development and Application of ML Model