A neural network model informs the total synthesis of clovane sesquiterpenoids

Efficient syntheses of complex small molecules, such as bioactive natural products, often involve detailed retrosynthetic planning and experimental evaluation of speculative synthetic routes. The central challenge of such an approach is that experimental evaluation of high-risk strategies is resource intensive because it requires iterative attempts at unsuccessful strategies. Along with the rapid development of cheminformatics and artificial intelligence, computer-aided synthetic planning has emerged to address this challenge. Herein, we report a complementary strategy that combines human-generated synthetic plans with computational prediction of the feasibility of key steps in the proposed synthesis. A neural network model (NNET) was trained on a literature-based dataset (from Reaxys) to predict the outcome of a generally disfavoured transformation, 6-endo-trig radical cyclization. The model performance was rigorously tested by experimental validation. On the basis of the virtual screening of potential substrates with our NNET model, optimal disconnections and structural modifications were chosen, resulting in five- to eight-step syntheses of three clovane sesquiterpenoids. This work establishes how a machine learning model informs human design and guides multistep syntheses of complex small molecules.

The synthesis of small molecules is integral to a variety of disciplines, from materials science to molecular devices to medicinal chemistry. For complex small molecules, efficient chemical synthesis requires detailed retrosynthetic planning and experimental evaluation. These plans usually involve one or more key steps that generate remarkable 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 because nuanced changes in substrate structure often result in notable changes in chemical reactivity that are challenging to predict.

One exciting approach to address the challenges associated with synthetic design is computer-aided synthetic planning, wherein computational approaches are used to provide synthetic routes. However, creative human-designed plans are valuable and crucial, especially in the context of highly complex small molecules. 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 (NNET) that is used to evaluate human-generated synthetic strategies towards clovane sesquiterpenoids by predicting the yields of key 6-endo-trig radical cyclization steps only on the basis of chemical structures (Fig. 1a). Efficient iterative virtual screening enabled us to choose ideal synthetic routes for multiple targets, which demonstrates the successful application of a machine learning (ML) model to guide target-oriented synthesis. Moreover, the success of this strategy argues for broader use of computational tools as part of the process for synthetic planning, and through this human–computer collaboration we highlight how human and computer planning need not be at odds.
Complex molecule synthesis has been limited because it is difficult to predict their outcomes. Baldwin's and Beckwith's rules 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 ML model described herein, which removes the obstacle of using the often unfavourable 6-endo-trig radical cyclization as the key disconnection for the synthesis of complex molecules.

To develop and apply an ML model to complex molecule synthesis, we devised the following workflow (Fig. 1b): (1) a library of literature examples from Reaxys was collected and annotated with chemical descriptors from simple and readily conducted density functional theory (DFT) calculations; (2) different ML model architectures were trained and evaluated for predictive performance; (3) human-generated retrosynthetic disconnections were evaluated using the trained models.

**Fig. 1 | ML model informs synthetic plan for clovane sesquiterpenoids.**

a, Clovane-type natural product synthesis is proposed via a 6-endo-trig radical cyclization to form the B ring, but the feasibility and optimal substrate are uncertain. (A, B and C labels in compound 1 are used to highlight specific rings). b, Workflow for the development and application of a ML model to guide synthetic planning. Bz, benzoyl; SM, starting material; TG, target.

### Database mining and parameter calculation
- Source literature examples from Reaxys
- Calculate chemical descriptors using DFT

### Model development and validation
- Train models based on different preprocessing methods and algorithms

### Evaluation of human-generated routes
- Predict yield for key disconnections
- Select the most promising route

### Substrate refinement
- Screen structural modifications for key intermediate and synthesize in laboratory

**Potential disconnections**

| Precursor | Predicted yields |
|-----------|------------------|
| a         | 10%              |
| b         | 50%              |
| c         | 80%              |
| ...       | ...              |

**Optimal structural modifications**

- Substrates
- Predicted yields
- Select the most promising route

**Fig. 1** ML model informs synthetic plan for clovane sesquiterpenoids. a, Clovane-type natural product synthesis is proposed via a 6-endo-trig radical cyclization to form the B ring, but the feasibility and optimal substrate are uncertain. (A, B and C labels in compound 1 are used to highlight specific rings). b, Workflow for the development and application of a ML model to guide synthetic planning. Bz, benzoyl; SM, starting material; TG, target.
ML model and (4) for the selected disconnection, substituents and functional groups were virtually screened with the model.

The feasibility of using ML to enable the total synthesis of clovanes is supported by complementary research in synthetic methods development using chemoinformatics21–24. These workflows inspired our efforts, but none of them could be directly applied to complex molecule synthesis. The major differences are summarized here: (1) the substrates used in synthetic methodology development are readily available for experimental screening and high-throughput experimentation, whereas substrates involved in complex molecule synthesis require time consuming multistep synthetic operations to obtain; (2) 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 and (3) the datasets generated from a single source (such as high-throughput experimentation) or a small number of literature references are relatively homogenous, whereas datasets derived from highly heterogenous sources possess potentially challenging variability25. The success of our effort demonstrates that a synthetically useful model can be developed by carefully selecting reliable data from search engines (such as Reaxys and SciFinder) without resorting to experimentally generating new datasets.

Although a purely DFT approach was successful for substrate selection in the case of the total synthesis of paspaline A and emindole PB26, methods that evaluate energies of multiple intermediates and transition states would be challenging for this radical cyclization if the entire pathway needed evaluation. It is generally assumed that the 5-exo mode of cyclization is kinetically favoured whereas the desired 6-endo mode of cyclization is thermodynamically favoured (Fig. 2a). Therefore, a rapid calculation would be to examine the ground state energies after cyclization and optimize for the thermodynamic favourability of the 6-endo cyclization. However, it was unknown whether greater thermodynamic preference (ΔG°f) would result in higher yield of the 6-endo product27. To investigate this possibility, the experimental yields of more than 100 literature reactions were plotted against their computed free energies of reaction (ΔG°f) in Fig. 2b. The lack of a correlation suggests that yield is determined by many factors in addition to ΔG°f. It was thus proposed 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 C(sp3)-centred radicals undergoing intra-molecular 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 descriptors28. A total of 340 descriptors per reaction were extracted to constitute the input parameters, including molecular, atomic, steric descriptors and linear combinations (Supplementary Table 4). Next, the library was split into training and test datasets (70/30) by the Kennard–Stone sampling to guarantee that the maximal breadth of feature space is covered in the training data29. As a large number of descriptors (340) were used relative to the small library size (99), overfitting was a major concern. Therefore, feature selection with correlation filtering (cut-off of 0.90) and dimensionality reduction with principal component analysis (threshold of 0.90)29 were used to transform 340 descriptors into 20 parameters.

An array of supervised ML models was tuned with tenfold cross-validation on training data and then were evaluated against the test dataset to provide R2 and MAE (mean absolute error) values. As shown in Fig. 2c, SIMPLS (statistically inspired modification of the partial least squares) and kNN (k-nearest neighbours) algorithms showed moderate predictive performance on the test dataset with R2 values of 0.56 and 0.59, respectively. A random forest model provided better performance with R2 = 0.79. A single hidden layer NNET delivered improved over these methods, providing an R2 value of 0.82, with an MAE of 12.1%. While using one or a few chemical descriptors could not allow for useful predictions to be made, the use of many more features did allow for useful predictions. This may be a function of the underlying importance of many possible factors that determine the efficiency of such radical cyclizations.

To evaluate the soundness of our NNET model, tenfold cross-validation and leave-one-out cross-validation (LOO-CV) were conducted on the whole library, providing slightly higher mean errors of 14.2% and 14.4%, respectively: the decreased Q2,LOO-CV (0.59) may be an indicator of overfitting, which prompted a need to further evaluate the model of the library with our NNET model. For this reason, an experimental validation study was conducted and is described at the end of the paper. Two additional control experiments were conducted (Fig. 2d): Y-randomization, in which yields are randomly shuffled across the training dataset and a random data test, in which chemically meaningful descriptors are replaced with randomly generated values30. The low correlations observed (R2 = 0.02 and 0.01, respectively) suggest that the predictions of our NNET model were achieved by identifying relationships between yield and chemically meaningful featureization, rather than by finding chance correlations. To test the model’s ability to extrapolate beyond the template library28, literature validation was conducted with an additional 26 examples of 6-endo radical cyclization from Reaxys and SciFinder; these substrates contained special functional groups that are not represented in the training or testing datasets, such as -CF3, substitution or heteroatoms (N, O) within the formed six-membered ring. We were pleased to find that reasonable correlation was observed, even though these key intermediates do not lie within the chemical space covered by the training data (Supplementary Fig. 13). The lower correlations (R2 = 0.63) and higher MAE (15.7%) indicate the limitations of our model, but even for this alternative substrate type, the degree of correlation could be useful in some contexts. For the purposes of clovane sesquiterpenoid synthesis, which have an all-carbon skeleton, it was not necessary to have high performance for these substrate types. Moreover, the reasonable performance of extrapolation further suggests the model identifies chemically meaningful information from physics-based features.

With the trained NNET model, different disconnections of the B ring corresponding to different synthetic routes to clovan-2,9-dione (1) were evaluated (Fig. 3a). The predicted yields of 6-endo-trig radical cyclizations from precursors 7, 8 and 9 are 26%, 46% and 34%, respectively. Due to limited available precedent for cyclizations of this type21–24, conventional logic would have discouraged those disconnections and 5-exo products would have been anticipated, but the model’s encouraging predictions for 8 mitigated that concern. Conventional disconnections to favour the 6-endo cyclization, such as the use of an enone instead of an alkene as the acceptor, did not provide significantly higher yield predictions (Supplementary Fig. 10). Ultimately, precursor 8 was selected, as it has a synthetically useful predicted yield and represents an innovative disconnection that leads to greater synthetic accessibility and more ready diversification to a variety of clovanes.

The next consideration investigated which proximal and remote functionality would be the optimal choice for the substrate given synthetic accessibility, predicted efficiency and use in accessing a variety of clovanes. A selection of substrates (10–14), which would readily lead to other clovane natural products, from more than 100 predictions (Supplementary Fig. 11) is shown in Fig. 3a to illustrate the planning considerations that were made. For example, the introduction of an additional carbonyl group in triketone 11 has a higher predicted yield that is qualitatively in line with expert intuition. Meanwhile, other modifications at sites distal to the reaction site lead to limited variability and uniformly synthetically useful yields are predicted.

As shown in Fig. 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 (15). The vicinal difunctionalization with a vinyl cuprate nucleophile and enolate trapping with HC(O)Me3 provided 16. Adduct 16 underwent a Robinson annulation with ethyl vinyl ketone to afford 17. An enone-selective Pd-catalysed hydrosilylation\(^{32}\) of 17 provided 18. These newly developed conditions were necessary as Pt, Cu and Rh catalysts provided inferior results. Treatment of the enoxysilane 18 with PhSeCl provided the radical precursor 19 as an inconsequential mixture of diastereomers.

After a series of optimization of reaction conditions (for example, temperature, concentration, solvent and reaction time), the 6-endo radical cyclization of 19 was realized with the highest yield of 45%, providing clovan-2,9-dione (1) and 5-exo product in a ratio of 1:1. 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, we assume that the predicted yields represent optimal yields as well. The successful realization of this radical cyclization experiments and extrapolation for the optimal NNET model. \(R^2\), coefficient of determination; SIMPLS, statistically inspired modification of the partial least squares; kNN, k-nearest neighbours; RF, random forest; CV, cross-validation and LOO-CV, leave-one-out cross-validation.

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-exo-trig radical cyclization (\(\Delta G_{\text{react}}\), uB3LYP/6-31g(d)) do not correlate with cyclization yields. c, Performance of different ML algorithms on the test dataset for the yield predictions of 6-exo-trig radical cyclization. d, Control
resulted in a more efficient five-step synthesis of 1, compared to the previously disclosed 15-step racemic strategy. In addition, enantioenriched 17 could easily be prepared through a Corey–Bakshi–Shibata reduction (Fig. 4a) and re-oxidation sequence, leading to an eight-step asymmetric synthesis (previously completed in 17 steps).

As shown in Fig. 3a, the feasibility of radical cyclization of 10 was evaluated by our NNET model with a predicted yield of 51%. The experimental success of this transformation (from 23 to 24, Fig. 4a) enabled the first total syntheses of rumpbellclovane A (26) and canangaterpene II (2) in eight steps from commercially available 15. The key elements of the synthesis are selective reduction of 17 and a late-stage Baeyer–Villiger followed by selective transesterification (24 to 26). The structure of canangaterpene II (2) was revised from the previously proposed structure on the basis of biosynthetic considerations, nuclear magnetic resonance spectroscopy calculations and our synthesis of the revised structure (see Supplementary Information for details).

To rigorously test the model performance, we examined an additional seven radical precursors as an experimental validation. Counterintuitive 6-endo-trig radical cyclization enables the five-step total synthesis of clovan-2,9-dione (1).

In summary, this report describes a platform that combines creative human-generated synthetic plans with robust computational analysis for a challenging key step. ML models are trained from readily accessible literature examples (from Reaxys and SciFinder) to pre-evaluate model performance in organic synthesis can be variable, yet experimental error is not evaluated or reported. We suggest researchers adopt the practice of including errors to assist future data science efforts.

Moreover, with the model reported herein, dozens of substrates can be evaluated in one day, whereas accurate DFT calculations of the full pathway for more than 100 substrates would be computationally intractable for practical time scales, as a single substrate may require weeks. The substantial time investment required to conduct DFT calculations poses an obstacle to incorporating calculations in synthetic planning; the ability to rapidly apply a ML-based model foreshadows broader future use of such computational tools in synthetic planning.

In summary, this report describes a platform that combines creative human-generated synthetic plans with robust computational analysis for a challenging key step. ML models are trained from readily accessible literature examples (from Reaxys and SciFinder) to predict the yields of a generally disfavoured chemical transformation (6-endo-trig radical cyclization). An NNET model was used to guide the retrosynthetic analysis of several sesquiterpenoid natural products, resulting in 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.

**Methods**

Model development and testing was carried out in the R Studio (v.1.2.5001) integrated development environment (Supplementary Information). DFT calculations were performed using Gaussian 09/16 system on the Grace Cluster of the Yale University High-Performance Computing Center (Supplementary Information). The procedures for reactions are included in Supplementary Information.

The general procedure for the radical cyclization was as follows. To a flame-dried 50 ml three-necked, round-bottom flask equipped with a magnetic stir bar and a reflux condenser was added selenide substrate (1.0 equiv.) in o-xylene (15 ml, 0.005 M). The reaction vial was moved to a preheated oil bath, and stirred at 140 °C for 10 min. A solution of ACHN (1,1′-azobis-1-cyclohexanenitrile, 0.1 equiv.) and Bu3SnH (1.2 equiv.) in o-xylene (2 ml) was added via syringe pump over 6 h and stirred at 140 °C for an additional 6 h. The reaction mixture was cooled to room temperature and concentrated under reduced pressure by rotary evaporation to provide a tan oil. Purification by flash column chromatography on silica gel afforded the desired products.

**Data availability**
The data supporting the findings of this study are available within the paper and its Supplementary Information.

**Code availability**
All code used to support the findings of this work is supplied as Supplementary Information. The code is also available on GitHub (https://github.com/Newhouse-Group/6-Endo-Radical-Cyclization). Source data are provided with this paper.

**Online content**
Any methods, additional references, Nature Portfolio reporting summaries, source data, extended data, supplementary information, acknowledgements, peer review information; details of author contributions and competing interests; and statements of data and code availability are available at https://doi.org/10.1038/s44160-023-00271-0.
References

1. Corey, E. J. Robert Robinson Lecture. Retrosynthetic thinking—essentials and examples. Chem. Soc. Rev. 17, 111–133 (1988).
2. Shen, Y. et al. Automation and computer-assisted planning for chemical synthesis. Nat. Rev. Methods Primers 1, 23 (2021).
3. Coley, C. W., Eyke, N. S. & Jensen, K. F. Autonomous discovery in the chemical sciences part I: progress. Angew. Chem. Int. Ed. 59, 22858–22893 (2020).
4. Segler, M. H. S., Preuss, M. & Waller, M. Planning chemical syntheses with deep neural networks and symbolic AI. Nature 555, 604–610 (2018).
5. Szymkuc, S. et al. Computer-assisted synthetic planning: the end of the beginning. Angew. Chem. Int. Ed. 55, 5904–5937 (2016).
6. Marsh, C. J. et al. Network-analysis-guided synthesis of weisacoinine D and iljestrandinine. Nature 528, 493–498 (2015).
7. Mo, Y. et al. Evaluating and clustering retrosynthesis pathways with learned strategy. Chem. Sci. 12, 1469–1476 (2021).
8. Mikulak-Klucznik, B. et al. Computational planning of the synthesis of complex natural products. Nature 588, 83–88 (2020).
9. Bideau, F. L. et al. Tricyclic sesquiterpenes from marine origin. Chem. Rev. 117, 6110–6159 (2017).

10. Chung, H.-M. et al. Rumphellclovane B, a novel clovane analogue from the gorgonian coral rumphella antipathies. Bull. Chem. Soc. Jpn 84, 119–121 (2011).
11. Chung, H.-M. et al. Rumphellclovanes C–E, new clovane-type sesquiterpenoids from the gorgonian coral Rumphella antipathies. Tetrahedron Lett. 69, 2740–2744 (2013).
12. Cheng, X. et al. Neuronal growth promoting sesquiterpenes–neolignans; syntheses and biological studies. Org. Biomol. Chem. 10, 383–393 (2012).
13. Yang, J. et al. Asymmetric total synthesis of (−)-clovan-2,9-dione using Rhl(-)-catalyzed [3+2+1] cycloaddition of 1-yne-vinylcyclopropane and CO. Org. Lett. 19, 6040–6043 (2017).
14. Liu, G., Zhang, Z., Fu, S. & Liu, B. Asymmetric total synthesis of rumphellclovane E. Org. Lett. 23, 290–295 (2021).
15. de Souza, G. G. et al. Biotransformation of clovane derivatives. Whole cell fungi mediated domino synthesis of rumphellclovane A. Org. Biomol. Chem. 10, 3315–3320 (2012).
16. Wright, P. M., Seiple, I. B. & Myers, A. G. The evolving role of chemical synthesis in antibiotic drug discovery. Angew. Chem. Int. Ed. 53, 8840–8869 (2014).
17. Romero, K. J., Galliher, M. S., Pratt, D. A. & Stephenson, C. R. J. Radicals in natural product synthesis. Chem. Soc. Rev. 47, 7851–7866 (2018).
18. Sha, C.-K., Jean, T.-S. & Wang, D.-C. Intramolecular radical cyclization of silylacetylenic or olefinic α-iodo ketones: application to the total synthesis of (±)-modhephene. Tetrahedron Lett. 31, 3745–3748 (1990).
19. Baldwin, J. E. Rules for ring closure. J. Chem. Soc., Chem. Commun. 18, 734–736 (1976).
20. Jorner, K. et al. Organic reactivity from mechanism to machine learning. Nat. Rev. Chem. 5, 240–255 (2021).
21. Reid, J. P. & Sigman, M. S. Holistic prediction of enantioselectivity in asymmetric catalysis. Nature 571, 343–348 (2019).
22. Ahneman, D. T. et al. Predicting reaction performance in C–N cross-coupling using machine learning. Science 360, 186–190 (2018).
23. Shields, B. J. et al. Bayesian reaction optimization as a tool for chemical synthesis. Nature 590, 89–96 (2021).
24. Zahrt, A. F. et al. Prediction of higher-selectivity catalysts by computer-driven workflow and machine learning. Science 363, eaau5631 (2019).
25. Schwaller, P., Vaucher, A., Laino, T. & Reynold, J.-L. Prediction of chemical reaction yields using deep learning. Mach. Learn.: Sci. Technol. 2, 015016 (2021).

Nature Synthesis | Volume 2 | June 2023 | 527–534

Acknowledgements

We are grateful for financial support from Yale University, the Sloan Foundation, Boehringer Ingelheim, Genentech and the National Institutes of Health (grant no. R0100045). Student support included Chemical Biology Training grant (no. T32 GM067543 to R.L.C.) and an Anderson Postdoctoral Fellowship (P.Z.). We gratefully acknowledge Yale University’s High-Performance Computing Center for providing resources for this work. F. Menges of Yale Chemical and Biophysical Instrumentation Center is gratefully acknowledged for obtaining the high-resolution mass spectrometry data.

Author contributions

M.E., P.Z. and T.R.N. initiated the project. Y.Z. and J.E. synthesized clovan-2,9-dione. P.Z. and R.L.C. synthesized rumphellclovane A and canangaterpene II. J.E. and R.L.C. carried out DFT and nuclear magnetic resonance spectroscopy calculations. P.Z., M.E. and Y.Z. performed the ML modelling. P.Z. and J.E. carried out experimental validation. All co-authors wrote and edited the manuscript.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s44160-023-00271-0.

Correspondence and requests for materials should be addressed to Timothy R. Newhouse.

Peer review information Nature Synthesis thanks the anonymous reviewers for their contribution to the peer review of this work.
