Machine learning study of the molecular drivers of natural product prices

Vincent Blay†, Institute of Integrative Systems Biology (I²Sysbio), Universitat de València and Spanish Research Council (CSIC), Valencia, Spain; Department of Microbiology and Environmental Toxicology, University of California Santa Cruz, Santa Cruz, CA, USA
Jie Dong†, Xiangya School of Pharmaceutical Sciences, Central South University, Changsha, China
Andrés Moya †, Institute of Integrative Systems Biology (I²Sysbio), Universitat de València and Spanish Research Council (CSIC), Valencia, Spain; Foundation for the Promotion of Sanitary and Biomedical Research of Valencian Community (FISABIO), Valencia, Spain; CIBER in Epidemiology and Public Health (CIBEResp), Madrid, Spain

Received June 04 2021; Revised July 22 2021; Accepted August 02 2021; View online August 27, 2021 at Wiley Online Library (wileyonlinelibrary.com); DOI: 10.1002/bbb.2281; Biofuels, Bioprod. Bioref. 15:1820–1834 (2021)

Abstract: The price of chemicals is a very complex variable. It can be impacted by production costs but also by market and managerial factors, which may have complex relationships with molecular characteristics and the state of technology and society. In this work, we explore the extent to which molecular characteristics can help explain natural product prices with the aid of machine learning tools. We interpret models trained on molecular descriptors and molecular fingerprints. These models can explain a notable proportion of the variation in prices, suggesting that production and separation costs are a major contributor to current natural product prices. Some molecular properties stand out as key price drivers across the chemical space, including hydrophobicity and the presence of certain heteroatoms. On the other hand, we demonstrate how the application of cliff analysis to prices allows the identification of small chemical transformations that have a remarkable impact on prices. Overall, the work suggests that machine learning could help achieve more consistent and fairer pricing and provides specific examples of chemical transformations in which synthetic biology could add significant value. © 2021 Society of Chemical Industry and John Wiley & Sons, Ltd

Supporting information may be found in the online version of this article.

Key words: pricing; cheminformatics; natural products; plant products; bacterial products; commercialization

Introduction

Natural products are chemical compounds that are discovered in living organisms in nature. These compounds can be useful to a wealth of industries, including cosmetics, flavors and fragrances, phytochemicals and pharmaceuticals. In the pharmaceutical domain, natural products are often referred to as biologics. Commercially, natural products may not just be isolated from natural sources, but they are often produced...
by suitable production organisms or synthesized by chemical means.\(^9\) In this sense, synthetic biology allows more efficient production hosts to be engineered and even allows molecular structures and properties to be modified to provide superior value to consumers.\(^10\)

A variety of pricing strategies are in use in the chemical and biosynthetic industries. In some cases, prices are adjusted based on the estimated value that products provide to consumers.\(^11\) In the case of novel specialty compounds, costs likely have a strong impact on prices. In this sense, production and separation costs are key contributors to the prices of chemicals,\(^12\) with production costs being the bottleneck for the commercialization of many biomolecules.\(^13–15\)

Natural products can be highly complex molecules, and machine learning is proving very successful at navigating and identifying patterns in high-dimensional datasets.\(^16\) In the case of natural products and synthetic biology, it is helping design new synthetic pathways, discover new bioreactions, engineer biomolecules, predict biomolecular interactions, and understand multi-omics relationships, among others.\(^5,17–19\)

Models constructed with machine learning algorithms can not only be used to predict an endpoint, but they may also help understand key factors affecting those endpoints. Usually, models built with simpler algorithms offer better interpretability at the expense of weaker performance.\(^20\) In this study, interpretability is prioritized over predictive accuracy. The models used make use of molecular descriptors and molecular fingerprints to represent molecules.

Molecular descriptors are cheminformatic tools that assign numbers to molecules. These numbers may capture some aspects of the molecular structure and enable the study of molecules using machine learning or other types of mathematical models. A well-established application of molecular descriptors is exploring structure–activity relationships of medicinal compounds.\(^21\) More recently, molecular descriptors have also been used to analyze the space of natural products and suggest new lines of research.\(^22,23\) Thousands of molecular descriptors have been proposed over the years, and these can be computed using tools like ChemDes,\(^24\) PyBioMed,\(^25\) Mordred,\(^26\) Dragon, or PaDEL.\(^27\) Molecular Operating Environment (MOE) is powerful design software providing commonly used descriptors with more explicit information.

Molecular fingerprints are an alternative method to represent molecules numerically. They capture a kind of molecular profile in the form of bit or count vectors. In structural fingerprints, the vector elements or keys indicate the presence or frequency of certain chemical patterns in the molecule.\(^28\) A variety of fingerprints have been proposed. For instance, MACCS (Molecular ACCess System, also known as MDL keyset) is a class of structural fingerprints with 166 bits.\(^29\) Fingerprints can also be calculated by the above tools, and have been used for a variety of molecular modeling tasks.\(^30,31\)

The objective of this study is to explore and gain insights on the relationship between price and chemical structure of natural products. This is interesting for several reasons. On the one hand, the price of biomolecules is a very complex property and few studies have addressed its prediction, less so its interpretation. On the other hand, a better understanding of prices could enable fairer and more consistent pricing practices both for existing and new products, and it could also provide valuable insights about synthetic targets worth investigating. To this end, we train a variety of machine learning models on two different types of predictors, molecular descriptors and molecular fingerprints, and interpret the strongest, most consistent associations between price and individual predictors. Finally, we also demonstrate the possible application of activity cliff analysis to the prices of natural products.

**Materials and methods**

**Library preparation**

A diagram of the workflow followed in the work is depicted in Fig. 1. 1654 molecules were indicated in the BioAustralis (BA) structural catalog (version BioAustralis_April2020.sdf). From these, through some automated and manual curation, a total of 985 molecules could be assigned prices according to the vendor's price catalog and were used for further modeling. Prices were converted from AUD to USD at the rate 1 AUD = 0.77 USD. Most molecules were offered in two amounts and the largest one was considered to compute the specific price of the compound (USD/mg). Eight hundred forty molecules were retrieved from the Extrasynthese (ES) structural catalog (version Extrasynthese-Formule2021dollaravURL.sdf), all of which had USD prices assigned. These were also expressed on a weight basis (USD/mg). A total of 1825 molecules were thus used to study and model compound pricing. Finally, prices were log-transformed for further analysis.

**Molecular descriptors and fingerprints**

We used MOE to calculate 436 2D and 3D descriptors. They belong to a variety of classes, including physical property and charge-related, constitutional, connectivity, topological, pharmacophore feature, energy-related, and geometric descriptors. Given the large number of molecular descriptors...
available and the fact that they may contain overlapping information, we first conduct a feature selection step. First, the descriptors computed were filtered out if (1) they had constant or near constant values across all molecules, (2) they had a standard deviation less than 0.01, (3) they had at least one missing value, or (4) they had a Pearson correlation to any other computed feature higher than 0.95, in which case one of the two variables was dropped randomly. This resulted in a total of 411 2D features being available for modeling. Second, we included a dummy variable indicating the supplier of the compounds (SourceBA = 1 indicating BioAustralis vendor; SourceBA = 0 indicating Extrasynthese vendor). Then, the caret package from R language was used for feature selection. Recursive feature elimination (RFE) with 10-fold cross validation was used to process the selection.

On the other hand, we used PyBioMed to calculate the bit version of MACCS, ECFP4, and PubChem fingerprints, which have been widely applied in the drug discovery field, show good performance, and provide quite interpretable structural information. We also included the SourceBA dummy variable.

**Interpretable models**

To obtain interpretable models with acceptable accuracy, we compared 19 regression algorithms combined with three kinds of fingerprints and MOE descriptors mentioned above. The 19 algorithms were implemented in an integrated Python environment mainly based on *sklearn* software. First, all the models were built by using the default hyperparameters (if any), then the best selected models were tuned by using a randomized grid search of a pre-defined search space. In this project, we ran between 10 and 1000 iterations in the search space, until there was no obvious change in performance. We divided the dataset into training and test sets randomly with an 80:20 split. We identified the best performing models by considering figures of merit in tenfold cross-validation. These figures of merit include the mean absolute error (MAE), root mean square error (RMSE), and coefficient of determination ($R^2$). Their definitions are shown in Eqns (1)–(3), where $y_i$ represents the observed specific price for compound $i$, $\hat{y}_i$ is the corresponding price estimate by the model, and $m$ is the number of compounds considered in the set:

$$MAE = \frac{1}{m} \sum_{i=1}^{m} |y_i - \hat{y}_i|$$  \hspace{1cm} (1)

$$RMSE = \sqrt{\frac{1}{m} \sum_{i=1}^{m} (y_i - \hat{y}_i)^2}$$  \hspace{1cm} (2)

$$R^2 = 1 - \frac{\sum_{i=1}^{m} (y_i - \hat{y}_i)^2}{\sum_{i=1}^{m} (y_i - \bar{y})^2}$$  \hspace{1cm} (3)
\[ R^2 = 1 - \frac{\sum(y_i - \hat{y}_i)^2}{\sum(y_i - \bar{y}_{\text{mean}})^2} \]  

(3)

**Feature relevance analysis**

After identifying the most informative models, we focused on studying feature relevance. To this end, we used two different metrics: feature importance scores and SHAP (SHapley Additive exPlanations) values. On the one hand, importance scores quantify how much each input feature (i.e., each key or bit in the molecular fingerprint or each molecular descriptor) is used by the individual trees in the ensemble to make the prediction (it also takes into account how deeply in the tree the feature is being used). On the other hand, SHAP is a Python package developed for model interpretation and can be applied to many machine-learning models.\(^{33}\) It regards all features as ‘contributors’. For each prediction sample, the model generates a prediction value, and a SHAP value is assigned to each feature. Features with large SHAP absolute values are more important. One big advantage of the SHAP value is that it reflects the influence of a feature in the predictions for a specific sample and whether that influence is positive or negative.\(^{34}\) The SHAP values reported in the work correspond to the test set of molecules, although virtually equivalent results were obtained when using the training set. Both of tree-based feature importance and SHAP methods have been applied successfully in different projects to aid in the interpretation of machine-learning models.\(^{24,35}\)

**SALI analysis**

An activity cliff or structure–activity landscape index (SALI) analysis was conducted using OSIRIS DataWarrior 5.2.1.\(^{36}\) Molecules were loaded and pairwise similarities were computed using the SkelSpheres method, which computes the overlap between fingerprints built by hashing circular spheres of atoms and bonds onto 1024-bit vectors.\(^{37}\) Once the similarities were computed, pairs of molecules with similarity above 0.95 were linked pairwise, and a SALI index value was computed for each pair (Eqn (4)):

\[
\text{SALI}(m_1, m_2) = \frac{|\text{Price}(m_1) - \text{Price}(m_2)|}{1 - S(m_1, m_2)}
\]

(4)

where \(m_1\) and \(m_2\) indicate two molecules, and \(S\) is a chemical similarity function, in this case based on the overlap between SkelSpheres fingerprints.

For visualization purposes, molecules were distributed on a 2D space while trying to preserve the relative distance or dissimilarity between molecules (i.e., \(1 - S(m_1, m_2)\)). This was accomplished by using a rubberbanding forcefield algorithm.

**Results and discussion**

**Molecular fingerprints**

The analyses conducted in this work are illustrated in Fig. 1. The fingerprint-based analysis in this work involves three main steps: (1) fingerprint comparison and selection, (2) model selection, and (3) consensus identification of the most important features. The results for the different models with the three types of fingerprints are presented in the supporting information (Tables S1–S3), with the best combinations summarized in Table 1. The \(R^2\) values for the models trained on MACCS fingerprints were higher than those with ECFP4 or PubChem fingerprints when considering the same algorithm, which suggests that MACCS fingerprints are the most informative ones for this task. The three best-performing models were random forest (RF), LightGBM, and CatBoost, with cross-validation \(R^2\) values of 0.55, 0.54, and 0.53,
respectively (see also Figs S1–S3). Their performance on the test set was comparable, with $R^2$ values between 0.504 and 0.522.

Random forests are a family of algorithms that train large numbers of individual regression trees (i.e., an ensemble or forest). Each individual tree is trained independently on a bootstrapped sample of the training dataset, and it is only given access to a random subset of the input features, which decorrelates the individual trees. Predictions are generated through a consensus among the many individual decision trees. On the other hand, CatBoost and LightGBM are implementations of gradient boosting on regression trees. In the end, predictions also have a contribution from each individual tree, and thus it is an ensemble method like random forests. In boosting, the idea is that individual models, in this case regression trees, are trained sequentially, based on the residuals left after fitting the preceding models. Moreover, the learning rate is slowed down through a shrinkage parameter. With boosting, good results can be obtained even with very short trees or stumps. 38

We calculated the SHAP values and feature importance scores for each model built using MACCS. Features identified as relevant and common across the different models were considered particularly relevant to the price of compounds.

Figure 2 presents some plots that illustrate the relevance of different features (fingerprint bits or keys) for one model, and how the top features compare across the three models.

![Figure 2](image_url)

**Figure 2.** (a) Feature importance scores in the RF model using MACCS fingerprints (top 10 features are shown). (b) Venn diagram of the top 10 features generated based on feature importance scores from different algorithms trained on MACCS fingerprints. (c) The SHAP value plots of the RF model using MACCS fingerprints (the top 20 features are shown). (d) Venn diagram of the top 20 features based on SHAP values from different algorithms trained on MACCS fingerprints.
Table 2. MACCS bits identified as particularly important to natural product price by different algorithms.

| Bit         | Method | Definition         | SMARTS                      |
|-------------|--------|--------------------|-----------------------------|
| MACCS130    | A, B   | # QH > 1           | ![SMARTS for # QH > 1]      |
| MACCS139    | A, B   | # O > 3            | ![SMARTS for # O > 3]       |
| MACCS144    | A, B   | # 6M ring > 1      | ![SMARTS for # 6M ring > 1] |
| MACCS124    | A, B   | # Aromatic Ring > 1| Not applicable              |
| MACCS75     | A, B   | # C=C(A)A          | ![SMARTS for # C=C(A)A]     |
| MACCS65     | B      | # CC(C)(C)A        | ![SMARTS for # CC(C)(C)A]   |
| MACCS49     | B      | # C=C(C)C          | ![SMARTS for # C=C(C)C]     |
| MACCS140    | B      | # CH3 > 2          | ![SMARTS for # CH3 > 2]     |
| MACCS100    | B      | # 8M Ring or larger| e.g.: ![SMARTS for 8M Ring] |
| MACCS128    | B      | # ACH2AACH2A       | ![SMARTS for # ACH2AACH2A]  |
| MACCS92     | B      | # QCH3             | ![SMARTS for # QCH3]        |
| MACCS136    | B      | # Heterocycle      | ![SMARTS for # Heterocycle] |

A: Feature importance scores. B: SHAP method. Details about the definition can be found in the ‘Fingerprints Library’ of ChemDes.

Figure 3. Visualization of five MACCS bits especially relevant to price based on both feature importance scores and the SHAP method.

selected. The MACCS bits identified as most relevant are collected in Table 2. The five common MACCS bits identified, both based on importance scores and SHAP values, are illustrated in Fig. 3. The common features obtained from the SHAP method are illustrated in Fig. S4. From Fig. 2 and Table 2, we see that 40% of the most relevant features identified by the feature importance score method are common across the three algorithms, while 50% common features are obtained using the SHAP method. Moreover, the most important predictors identified using feature importance scores nicely overlap with those identified using the SHAP method, confirming that these features are especially relevant to price. The six common features are SourceBA and MACCS keys 130, 139, 144, 124, and 75.

By analyzing the definitions of the relevant MACCS keys, some structural insights can be gained. First, the number and complexity of rings in molecules have an important impact on price. In particular, MACCS keys 124, 144, and 100 are all related to rings, and from the SHAP values in Fig. 2(c), we see that higher feature values will have a positive effect on the price. This is also in line with our initial expectations: the more complex the compounds are, the higher the cost of separation and extraction or processing may be. Second, the presence of heteroatoms may also have an important effect on price. For example, MACCS bits 130, 139, 92, and 136 highlight the importance of chemical bonds formed by oxygen atoms and other non-hydrocarbon atoms. From Fig. 2(c), we see that MACCS keys 130, 139, and 136 exert a positive effect on price, whereas MACCS92 has a negative effect. Third, there is also a strong relationship between price and the saturation degree of compounds. For example, MACCS bits 128 and 92 indicate that more saturated structures may lead to a lower price. This rule could be strengthened by the positive contribution of MACCS keys 75 and 49.

Conventional molecular descriptors

In this section, we aim to identify a set of molecular descriptors that are especially relevant to natural product prices. To this end, we adapt the strategy used above, as
summarized in Fig. 1. In short, we train a variety of machine learning models using a set of MOE descriptors as inputs, and we then leverage the best performing models to rank variables by importance. The features that the models agree that are most important, we consider them particularly relevant to the price of natural products and interpret them.

Given the large number of molecular descriptors available and the fact that they may contain overlapping information, we first conduct a feature selection step (see the materials and methods section above). This procedure returned a subset of 59 features (Fig. S5). Based on these features, a variety of algorithms were explored. Table 3 summarizes the best three performing algorithms (figures of merit for all models are shown in Table S4). These algorithms are RF, CatBoost, and Gradient Boosting Regressor (GBR), another implementation of boosted regression trees, which attain cross-validation $R^2$ values of 0.5722, 0.5632, and 0.5547, respectively. The test set $R^2$ values are between 0.5055 and 0.5314, and the corresponding residual plots are shown in Fig. S6. The SHAP method used in this work does not support GBR and thus we also considered the LightGBM model.

Figure 4 displays the modeling results and feature importance scores for the RF model (the corresponding plots for CatBoost and GBR models are shown in Figs S7 and S8). The eight common features based on feature importance across the three algorithms are SourceBA, vsurf_HB2, pmi1, GCUT_SLOGP_0, h_emd, Q_VSA_FNEG, BCUT_SLOGP_3, and Kier3. The 12 common descriptors identified as relevant based on SHAP scores across the three algorithms SourceBA, vsurf_HB2, pmi1, GCUT_PEOE_0, GCUT_SLOGP_3, b_double, vsurf_D8, h_emd, h_log_pbo, logS, Kier3, pmiX. Finally, the five common features identified as particularly important by both the feature importance and SHAP method across the three algorithms are SourceBA, vsurf_HB2, pmi1, h_emd, and Kier3.

From these results, we observe that SourceBA has a similar impact on price as in the MACCS-based models, with microbial-produced products tending to have a higher price than plant-produced products. Plants may be more cost-effective producers of certain metabolites and may require a less stringent control of growth conditions, at the expense of longer production times. In this sense, plant cell cultures and even cell-free cultures may be interesting options.

Although the interpretation of molecular descriptors in terms of known chemical structures and properties can be challenging, we also attempted to gain chemical insights by analyzing the most relevant molecular descriptors. First, vsurf_HB2 is the H-bond donor capacity based on volume, whereas h_emd is the sum of hydrogen bond donor strengths. These definitions suggest that the presence of heteroatoms with high electronegativity, such as N and O, may lead to higher prices, which agrees with the insights gained from the MACCS fingerprint analysis above. Secondly, pmi1 represents the first diagonal element of diagonalized moment of inertia tensor, whereas Kier3 is the third kappa shape index. There is no simple interpretation of these descriptors, although the number of rings and branches may affect their values through the corresponding molecular graphs. This would agree with the observation from MACCS key 65. Thirdly, GCUT_SLOGP_0, GCUT_SLOGP_3, BCUT_SLOGP_3, h_log_pbo, logS are all strongly related with the hydrophilicity and lipophilicity of molecules. The SHAP values indicate that molecules with higher lipophilicity (higher GCUT_SLOGP_0, GCUT_SLOGP_3, BCUT_SLOGP_3, h_log_pbo, and lower logS values) tend to command a higher price. This may be related to molecules with low solubility tending to be produced in lower titers. Finally, the descriptor b_double may also support the relationship between the price and the saturation degree observed above.

Some limitations of the study include the few suppliers for which high-quality data could be accessed and the fact that it represents a static snapshot of prices. More importantly, price is a very complex property that may depend on many factors, including technical factors, such as titer, yield, and productivity, as well as market factors and managerial decisions. One such market factor is supply and demand. It might contribute to explain part of the price difference between, say, epirubicin, a product in relatively low demand, and doxorubicin, for many suppliers may have it in stock. Note also that the factors above may be interrelated. For example, high demand may motivate the arrival of new

Table 3. Summary of the top-performing algorithms trained on selected MOE descriptors.

| Algorithm | Cross validation | Test |
|-----------|-----------------|------|
|           | $R^2$ | RMSE | MAE | $R^2$ | RMSE | MAE |
| RF        | 0.5722 | 0.5302 | 0.3910 | 0.5302 | 0.5595 | 0.4107 |
| CatBoost  | 0.5632 | 0.5357 | 0.3927 | 0.5314 | 0.5588 | 0.4059 |
| GBR       | 0.5547 | 0.5410 | 0.4038 | 0.5055 | 0.5741 | 0.4286 |
| LightGBM  | 0.5449 | 0.5467 | 0.3983 | 0.5077 | 0.5728 | 0.4174 |
suppliers to the market, fostering competition, as well as the development of more cost-effective methods for compound production at scale, thus affecting prices. By contrast, limited demand and supply may increase prices, as compounds may not be available in stock and need to be produced on demand.

In this sense, and despite not being the focus of the work, the models developed were able to explain a remarkable proportion of the observed variation in prices clearly by just considering molecular information. One key component of prices is production costs, which may be strongly affected by molecular characteristics. In the case of the bacterial metabolites, the main price inputs are difficulty in fermentation and yield (BioAustralis, personal communication). Separation costs are also known to be an important cost component in extracting compounds from plants.\textsuperscript{46}

It must also be noted that, besides the algorithms considered in this work (e.g., Table S1), there are many others that could be applied if the goal were to improve the predictive performance of the modeling. Indeed, the development of new machine learning models is a vibrant area of research. For example, neural networks are able to...
attain very high expressivity and can deal naturally with high-dimensional inputs and multiple outputs.\textsuperscript{47} Compared to random forests, neural networks tend to have many more parameters and thus shine in data-rich scenarios, often in combination with regularization methods to avoid overfitting. On the other hand, the interpretability of neural networks may be more challenging than that of tree-based models, although great progress is being made. Similarly, the fingerprints and molecular descriptors explored as features in this work represent only a subset of the many that have been proposed and continue to be investigated to improve molecular models.\textsuperscript{48,49} More generally, fingerprints and descriptors are only two possibilities for encoding molecules into numeric features in a way that can be processed by (machine learning) mathematical models. We consider that the development of more informative molecular features for large and/or chiral molecules would be particularly beneficial to modeling properties and designing superior bioproducts.

**SALI analysis**

In addition to the patterns discovered in the previous two sections, we were also interested in identifying small chemical changes that might have a substantial impact on price. To this end, we conducted a cliff or SALI analysis of prices.\textsuperscript{50} A SALI index for a pair of molecules quantifies how much price changes relative to the change in molecular structure (Eqn (4)).

The results of the study are shown in Fig. 5. There, each point represents a compound, which is colored based on its price, whereas the size of the point is scaled based on the SALI value. The distance between points is proportional to the denominator in the SALI equation (chemical distance or dissimilarity). Finally, the background is colored based on the supplier of the compounds (SourceBA predictor).

Overall, we observe that compounds in the plant-derived catalog tend to cluster together in one region of space (pink background), whereas bacterial metabolites seem more structurally diverse and scattered throughout the chemical

---

**Figure 5.** Structure–activity landscape index analysis of the price of natural products in this study. Molecules are clustered by chemical similarity and information is represented on price (node color), supplier (background color), and SALI value (node size), which captures the magnitude of the change in price relative to a small chemical change.
Table 4. Selected pairs of molecules having small chemical differences but high price differences identified from the SALI analysis.

| Pair | Molecule 1 | Molecule 2 | Price 1 | Price 2 |
|------|------------|------------|---------|---------|
| 1    | Doxorubicin hydrochloride | Epirubicin hydrochloride | $657.58/100 mg | $6.58/mg (BA) |
|      | $431.2/5 mg | $86.24/mg (BA) |
| 2    | Cyclosporin A | Cyclosporin H | $256.14/100 mg | $2.56/mg (BA) |
|      | $321.09/5 mg | $64.22/mg (BA) |
| 3    | Amycolatopsin A | Amycolatopsin B | $565.95/5 mg | $113.19/mg (BA) |
|      | $405.25/0.5 mg | $808.5/mg (BA) |
| 4    | Conglobatin C | Conglobatin B | $565.95/5 mg | $113.19/mg (BA) |
|      | $431.2/0.5 mg | $862.4/mg (BA) |
| 5    | (±/−)-linalool | (−)-linalool | $62/100 mg | $0.62/mg (ES) |
|      | $50/50 mg | £1.00/mg (ES) |
| 6    | (±/−)-camphor | (+)-camphor | $50/1 mg | $50/mg (ES) |
|      | $50/1 mg | $50/mg (ES) |
Table 4. (Continued)

|   | ![Molecular Structure](image1) | ![Molecular Structure](image2) |
|---|--------------------------------|--------------------------------|
|   | Maduramicin ammonium $307.23/25$ mg $12.29/mg$ (BA) | Maduramicin $447.37/25$ mg (BA) $17.89/mg$ (BA) |
| 8 | ![Molecular Structure](image3) | ![Molecular Structure](image4) |
|   | Valnemulin hydrochloride $242.55/100$ mg $2.43/mg$ (BA) | Valnemulin $307.23/25$ mg $12.29/mg$ (BA) |
| 9 | ![Molecular Structure](image5) | ![Molecular Structure](image6) |
|   | Thiostrepton $477.4/25$ mg $19.10/mg$ (BA) | Siomycin A $889.35/2.5$ mg $355.74/mg$ (BA) |
| 10 | ![Molecular Structure](image7) | ![Molecular Structure](image8) |
|   | Desertomycin A $706.09/5$ mg $141.22/mg$ (BA) | Desertomycin B (Oasomycin B) $404.25/0.5$ mg $808.50/mg$ (BA) |
| 11 | ![Molecular Structure](image9) | ![Molecular Structure](image10) |
|   | Carbazomycin C $565.95/5$ mg $113.19/mg$ (BA) | Carbazomycin D $404.25/0.5$ mg $808.50/mg$ (BA) |
bacterial products than to other plant-derived compounds. Some examples are vulpic acid, which closely resembles the bacterial vulpinic acid, or everinic acid, which is a methylated analog of the bacterial lecanoric acid. We also observe that, overall, the plant-derived compounds represented here tend to concentrate in a narrower band of relatively lower prices, whereas specific prices of microbial-derived compounds span over three orders of magnitude (see color bar).

The SALI analysis allows to identify interesting pairs of compounds in which one of the molecules in the pair has a surprisingly high or low price compared to the other. Table 4 summarizes some illustrative examples, which we comment on next:

1. In large compounds, small changes can make a big difference. For example, a change in stereochemistry (doxorubicin/epirubicin), the removal of a hydroxyl group, or the introduction of a methyl group can drive several-fold changes in the price of some molecules (Table 4, entries 1–4).
2. It may be favorable to develop methods to purify products enantiomerically. For example, the selling price of (−)-linalool is 60% higher than its racemate (Table 4, entry 5). Of course, this may depend on the specific compound. For example, enantiomerically pure camphor has a similar price to the racemate (Table 4, entry 6).
3. It may be favorable to price to remove salts from the product, as illustrated by the cases of maduramicin and vanemulin, for example (Table 4, entries 7 and 8).
4. The ability to control the sequence of building blocks that lead to the natural product can be particularly advantageous, as illustrated by the cases of siomycin A versus thiostrepton, or desertomycin A versus oasomycin B (Table 4, entries 9 and 10). In this sense, modular biosynthetic systems, such as polyketide synthases, hold great promise.
5. It may be interesting to apply small post-synthetic transformations of natural products. For example, the deamidation of venturicidin A could lead to the production of the more highly priced venturicidin B. Similarly, the partial hydrogenation of meleagrin could lead to neoxaline, sold at a tenfold higher price. The partial hydrogenation of 4′-acetylchrysomycin B would lead to 4′-acetylchrysomycin C, which was sold at a fourfold higher specific price. Processes like these...
(Table 4, entries 11–14) seem well suited for post-synthesis modification using abiotic catalysts.

Conclusion

In this work, we explored the application of machine learning to understand some important drivers of natural product prices. The results demonstrate that chemical structures alone can explain a considerable variation of the current prices in this sector. This may occur through complex, unobserved biotechnological variables, such as production and separation costs, yield, titers, and productivity. Valuable insights can be obtained at a high level across the dataset, which spans very diverse chemical families, but also at a very granular level by looking at small clusters of compounds.

By applying different strategies to build interpretable models, we successfully explain and visualize the impact of some key structural characteristics on prices. The analyses based on molecular descriptors and fingerprints provided insights that are highly consistent with each other. Overall, fingerprint-based models attained a slightly lower performance than molecular descriptor-based models, although they are more easily interpretable. Salient molecular features identified as relevant to price include lipophilicity, which may limit the yield, titers, or productivity attainable of a given product, and the presence of rings and heteroatoms, which may affect the difficulty of separation and the bioactivity of compounds, contributing to their cost and demand, respectively. Finally, we demonstrate the application of activity cliff analysis to compound prices. The findings highlight the interest in developing synthetic strategies that can deliver pure bioproducts with superior regio- and stereospecific control.

The rules identified can not only reveal the relationship between structure and price of natural products, which may inform better pricing strategies but also provide new ideas to chemists and synthetic biologists on potentially attractive biosynthesis targets.

Acknowledgements

The authors thank Christelle Merlet (Extrasynthese), Sandra Seymour (BioAustralis), and Phillipa Smith (Microbial Screening Technologies) for their help building the compound libraries in this work. This project is funded by Generalitat Valenciana (Spain) projects PROMETEO/2018/133 (to A.M.) and APOSTD/2020/120 (to V.B.), National Natural Science Foundation of China (Grant no. 22003078, to J.D.) and Hunan Provincial Natural Science Foundation of China (Grant no. 2019JJ51003, to J.D.).

References

1. NCB, All natural. Nat Chem Biol 3:351–351 (2007). https://doi.org/10.1038/ncmbio0707-351.
2. Shim J-H, Inhibitory effects of cyclohexamidin on melamin synthesis. Molecules 26:2526 (2021). https://doi.org/10.3390/molecules26092526.
3. Bomgardner, M.M., 2019. How perfumers walk the fine line between natural and synthetic. Chemical & Engineering News (vol. 97).
4. Sharmane JB, Mahomoodally FM, Zengin G and Maggi F, Essential oils as natural sources of fragrance compounds for cosmetics and cosmeceuticals. Molecules 26:666 (2021). https://doi.org/10.3390/molecules26030666.
5. Tran LT, Blay V, Luang S, Eurltong C, Choknud S, González-Díaz H et al., Engineering faster transglycosidases and their acceptor specificity. Green Chem 21:2823–2836 (2019). https://doi.org/10.1039/C9GC00621D.
6. Yin S, Mei Y, Wei L, Zou L, Cai Z, Wu N et al., Comparison of multiple bioactive constituents in the corolla and other parts of Abelmoschus manihot. Molecules 26:1864 (2021). https://doi.org/10.3390/molecules26071864.
7. Atanasov AG, Zotchew SB, Dirsch VM and Supuran CT, Natural products in drug discovery: Advances and opportunities. Nat Rev Drug Discov 20:200–216 (2021). https://doi.org/10.1038/s41573-020-00114-2.
8. Morrow T and Felcone LH, Defining the difference: What makes biologics unique. Biotechnol Health 1:24–29 (2004).
9. Pham JV, Yilma MA, Feliz A, Majid MT, Maffetone N, Walker JR et al., A review of the microbial production of bioactive natural products and biologics. Front Microbiol 10:1404 (2019). https://doi.org/10.3389/fmicb.2019.01404.
10. Meng F and Ellis T, The second decade of synthetic biology: 2010–2020. Nat Commun 11:5174 (2020). https://doi.org/10.1038/s41467-020-19092-2.
11. Burns, D., Schottland, D., 2018. The Formula for Better Pricing in Chemicals, [WWW Document]. Bain & Company. Available: https://www.bain.com/insights/the-formula-for-better-pricing-in-chemicals/ [5 October 2021].
12. Christopherson PC, Product pricing in the chemical industry. Financ Anal J 33:54–62 (1977).
13. Klein-Marcuschamer D, Oleskowicz-Popiel P, Simmons BA and Blanch HW, The challenge of enzyme cost in the production of lignocellulosic biofuels. Biotechnol Bioeng 109:1083–1087 (2012). https://doi.org/10.1002/bit.24370.
14. Rowe GE and Margaritis A, Bioprocess design and economic analysis for the commercial production of environmentally friendly bioinsecticides from Bacillus thuringiensis HD-1 kurstaki. Biotechnol Bioeng 86:377–388 (2004). https://doi.org/10.1002/bit.20146.
15. Wu W and Maravellias CT, Synthesis and techno-economic assessment of microbial-based processes for terpenes production. Biotechnol Biofuels 11:294 (2018). https://doi.org/10.1186/s13068-018-1285-7.
16. Lawson CE, Marti JM, Radiojovic T, Jonnalagadda SVR, Gentz R, Hillson NJ et al., Machine learning for metabolic engineering: A review. Metab Eng 63:34–60 (2021). https://doi.org/10.1016/j.menb.2020.10.005.
17. Silva JCF, Teixeira RM, Silva FF, Brommonschenkel SH and Fontes EPB, Machine learning approaches and their current application in plant molecular biology: A systematic review. Plant Sci 214:37–47 (2019). https://doi.org/10.1016/j.plantsci.2019.03.020.
18. Watanabe N, Ohnuki Y and Sakakibara Y, Deep learning integration of molecular and interactome data for protein–
compound interaction prediction. J Chem 13:36 (2021). https://doi.org/10.1186/s13321-021-00513-3.
19. Xu C and Jackson SA, Machine learning and complex biological data. Genome Biol 20:76 (2019). https://doi.org/10.1186/s13059-019-1689-0.
20. ElShawi R, Sherif Y, Al-Mallah M and Sakr S, Interpretability in healthcare: A comparative study of local machine learning interpretability techniques. Comput Intell 1–18 (2020). https://doi.org/10.1111/coin.12410.
21. Dudek AZ, Arodz T and Gálvez J, Computational methods in developing quantitative structure–activity relationships (QSAR): A review. Comb Chem High Throughput Screen 9:213–228 (2006). https://doi.org/10.2174/1386207076776055539.
22. Pye OR, Bertin MJ, Lokey RS, Gerwick WH and Lintingon RG, Retrospective analysis of natural products provides insights for future discovery trends. Proc Natl Acad Sci U S A 114:5601–5606 (2017). https://doi.org/10.1073/pnas.1614880114.
23. Stratton CF, Newman DJ and Tan DS, Cheminformatic comparison of approved drugs from natural product versus synthetic origins. Bioorg Med Chem Lett 25:4802–4807 (2015). https://doi.org/10.1016/j.bmcl.2015.07.014.
24. Dong J, Cao D-S, Miao H-Y, Liu S, Deng B-C, Yun Y-H et al., ChemDes: An integrated web-based platform for molecular descriptor and fingerprint computation. J Chem 76 (2015). https://doi.org/10.1016/j.sciint.2015-01-0109-z.
25. Dong J, Yao Z-J, Zhang L, Luo F, Lin Q, Lu A-P et al., PyBioMed: A python library for various molecular representations of chemicals, proteins and DNAs and their interactions. J Chem 10:16 (2018). https://doi.org/10.1186/s13321-018-0270-2.
26. Moriwaki H, Tian Y-S, Kawashita N and Takagi T, Mordred: A descriptor and fingerprint computation. Chimia 72:877–883 (2018). https://doi.org/10.1002/chmi.201706677.
27. Yu et al., The screening compound collection: A key asset for drug discovery. J Mol Struct (THEOCHEM) 1131:106–107 (2016). https://doi.org/10.1016/j.molstruc.2015-11-00610.
28. Berzin MA and Prado JM, Molecular drivers of natural product prices V Blay et al.
© 2021 The Authors. Biofuels, Bioproducts and Biorefining published by Society of Industrial Chemistry and John Wiley & Sons Ltd.

Biofuels, Bioprod. Bioref. 15:1820–1834 (2021); DOI: 10.1002/bbb.2281
Vincent Blay

Vincent Blay received a PhD in chemical engineering from the University of Valencia, Spain, and an MBA from the Ohio State University, USA. He has conducted research at several institutions worldwide, including the University of California, USA. His interests include studying and engineering molecular interactions for biomedical and synthetic biology applications, in combination with artificial intelligence methods.

Jie Dong

Jie Dong received a PhD from Central South University, China. He conducted postdoctoral research at the State Key Laboratory of Quality Research in Chinese Medicine, University of Macau. He is currently an associate professor at the Xiangya School of Pharmaceutical Sciences, Central South University, China. His interests include artificial intelligence systems and the development of software, web services, and databases for cheminformatics and drug discovery.

Andrés Moya

Andrés Moya has doctorates in biology and philosophy, and is a professor of genetics at the University of Valencia, Spain. His scientific and intellectual activity focuses on the fields of genetics and evolution, and philosophy. He investigates the human and animal microbiome using the integrative and systems biology approaches. He also carries out extensive work in dissemination and reflection on science.