ChemSusChem
Research Article
doi.org/10.1002/cssc.202300538
Chemistry Europe
European Chemical Societies Publishing
www.chemsuschem.org

Synthesis of Pyrrole-2-Carboxylic Acid from Cellulose- and Chitin-Based Feedstocks Discovered by the Automated Route Search

Thuy Trang Pham,[a] Zhen Guo,[b, c] Bing Li,[a] Alexei A. Lapkin,*[b, d] and Ning Yan*[a]

The shift towards sustainable feedstocks for platform chemicals requires new routes to access functional molecules that contain heteroatoms, but there are limited bio-derived feedstocks that lead to heteroatoms in platform chemicals. Combining renewable molecules of different origins could be a solution to optimize the use of atoms from renewable sources. However, the lack of retrosynthetic tools makes it challenging to examine the extensive reaction networks of various platform molecules focusing on multiple bio-based feedstocks. In this study, a protocol was developed to identify potential transformation pathways that allow for the use of feedstocks from different origins. By analyzing existing knowledge on chemical reactions in large databases, several promising synthetic routes were shortlisted, with the reaction of D-glucosamine and pyruvic acid being the most interesting to make pyrrole-2-carboxylic acid (PCA). The optimized synthetic conditions resulted in 50% yield of PCA, with insights gained from temperature variant NMR studies. The use of substrates obtained from two different bio-feedstock bases, namely cellulose and chitin, allowed for the establishment of a PCA-based chemical space.

Introduction

Reduction of carbon-intensity of chemistry-using industries can be achieved through use of bio-renewable feedstocks that can simultaneously provide access to low embedded fossil carbon feedstock molecules and useful molecular scaffolds.[1] Among various biomass sources, plant-based lignocellulose (cellulose, hemicellulose and lignin) and crustacean/insect/fungi-based chitin are particularly attractive, being by-products in various manufacturing processes. Multiple transformation pathways to convert these biopolymers into a broad range of chemicals had been studied, leading to the synthesis of valuable platform chemicals and further to downstream products.[2] Lignocellulose biomass is an ideal starting material to access C-, H-, and O-containing chemicals through key intermediates such as glucose,[2c,3] furfural,[4] 5-hydroxymethylfurfural,[5] and pyruvic acid (PA).[6] On the other hand, chitin, which contains biologically-fixed nitrogen, has been made into a range of organonitrogen compounds that can potentially be used to manufacture fabrics, pharmaceuticals, surfactants, and nutrients. The key platform chemicals identified so far include D-glucosamine ($GlcNH_2$), N-acetylglucosamine (NAG), and 3-acetamido-5-acetylfuran.[7] In these previous studies, conversions of lignocellulose and chitin were based on independent synthetic routes. We hypothesize that integrating the two platforms and identifying routes that are based on elements sourced from both biorefining platforms, would provide more diversified products, bearing structures that are difficult to build from a single source of biomass. While it represents an intriguing approach to make fossil carbon-free and Haber-Bosh-independent nitrogen-containing compounds, to our knowledge, little study has been conducted thus far in this area.

One reason for scarcity of such studies is the large number of possible transformations and synthetic routes that must be evaluated in developing the synthetic routes starting from very different raw materials and involving different intermediates. Conventional manual literature research can hardly figure out the impact of new molecules and reactions on the whole potential network of reactions. Such a task requires an automated approach, in which assembly and evaluation of potential transformations could be done algorithmically.

Amongst algorithmic chemical routes search techniques, typically grouped under the umbrella term Computer Assisted Synthesis Planning (CASP), the synthesis planners such as ASKCOS (developed by MIT and freely available online),[8] the Reaxys Retrosynthesis,[9] or IBM RXN[10] and MANIFOLD by PostEra,[11] are designed to suggest potential reaction sequences

---

[a] Dr. T. T. Pham,+ Dr. B. Li, Prof. N. Yan
Department of Chemical and Biomolecular Engineering
National University of Singapore
4 Engineering Drive 4, 117585 Singapore City (Singapore)
E-mail: ning.yan@nus.edu.sg

[b] Dr. Z. Guo,+ Prof. A. A. Lapkin
Cambridge Centre for Advanced Research and Education in Singapore (CARES Ltd)
1 CREATE Way, #05-05 Create Tower, 138602 Singapore City (Singapore)

[c] Dr. Z. Guo+
Chemical Data Intelligence (CDI) Pte Ltd
Robinson Road #02-00, 068898 Singapore City (Singapore)

[d] Prof. A. A. Lapkin
Department of Chemical Engineering and Biotechnology
University of Cambridge
Cambridge CB3 0AS (UK)

Supporting information for this article is available on the WWW under https://doi.org/10.1002/cssc.202300538

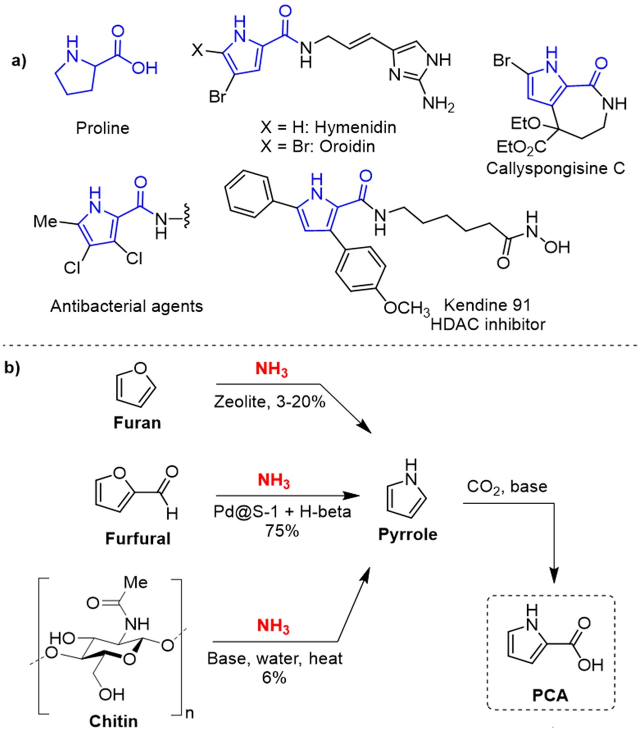
to a target compound typically from a specified starting molecule or a range of commercially available starting molecules. These tools are machine-learning based algorithms trained on various fractions of the available chemical data, converted into different machine-readable representations, and then interpreted using different learning algorithms.

An orthogonal algorithmic method is based on hard-coded rules of organic chemistry extracted from literature data by hand and then encoded into an expert system, for example the former Chematica,[12] now owned by Merck and marketed as Synthia. Such a tool may also work in parallel with access to large databases of organic reactions to verify the suggested routes.

The graph-based approaches give the third option of addressing the challenge of algorithmic search for potentially unique reactions in large networks.[13] The pioneering work of analysis reaction networks was done by Grzybowski et. al. They analyzed the evolution of a reaction network over time, and found that a reaction network is similar to that of World Wide Web, forming a scale-free network.[13a] The same group conducted searching of synthesis routes in bipartite networks with the guidance of an empirical synthesis cost function involving material and labour costs.[13e] Subsequently more scoring and cost functions were developed, which opened the era of practical application of synthesis planning techniques.[13c] In our group, we have shown that network searching can be used to identify and rank routes between new platform molecules and the desired target functional molecules,[14] or to identify the list of highly connected molecules as targets for synthesis from bio-feedstocks.[15] However, as Grzybowski highlighted in one of his papers titled "Computer-Assisted Synthetic Planning: The End of the Beginning", the automatic synthesis planning technique is still far from maturity, despite the significant progress of recent years.[16,17] Challenges include but not limited to: poor data quality (sparse and low integrity), high computation cost due to overwhelming synthesis routes (i.e., enormous combinations of reactions, known as "combinatorial explosion"), diverse searching objectives (costs, sustainability, safety etc.), complexity of de novo synthesis planning etc. Our own focus was on developing a synthetic planning tool for practical applications, especially development of sustainable synthetic routes.[18] A workflow consisting of mining reaction network and exergy optimization algorithms was developed to optimize synthetic design, and was exemplified on a case study of conversion of β-pinene to citral.[19] To address the issue of sparse data, we performed deep mining from reaction network down to molecular substructure and atom-level analysis. This was achieved by developing similarity-based heuristics and in-depth molecular structure analysis. The deep mining approach was applied to predict reaction impurities which are missing in most reaction records.[20] The concept of object-oriented searching was also introduced into the design of our searching tool, allowing users to customize their searches by utilizing a combination of various heuristics to meet their specific objective.

In this work we hypothesized that we should be able to identify reaction pathways of important chemical transformations in large reaction networks when searching for opportunities in using bio-feedstocks of different origin. A brief search showed that most reactions have very low yields, especially reactions reported long time ago. However, this does not necessarily imply that these reactions are obsolete or have no potential for improvement with current technologies. Thus, similarity-based heuristics, as metrics of usefulness of reactions, were utilized to evaluate reactions and routes. And normally we do not want to lose useful substructures of feedstocks, such as the five member rings of platform-molecules and nitrogen of chitin. Substructure analysis was also applied throughout most searches. As a proof of concept, we applied graph-based approaches to the synthetic planning of pyrrole-2-carboxylic acid (PCA), using cellulose and chitin as feedstocks, through datamining Reaxys®.‡ PCA is an important platform intermediate and building block for a number of high-value products, including amino acid proline, bioactive marine natural products as well as synthetic bioactive compounds (Scheme 1a). For instance, pyrrole-2-carboxamide is a common fragment present in marine sponges' alkaloids such as hymenidin, oroidin, and callyspongisine C,[21] whereas 3,4-dichloro-5-methyl-1H-pyrrole-2-carboxamide is a molecular fragment of synthetic antibacterial compounds, crucial for binding to the active site of bacterial topoisomerases.[22] Kendine 91, a histone deacetylase (HDAC) inhibitor, also possesses the pyrrole-2-carboxamide moiety.[23]

Although PCA is conventionally made from petroleum-based feedstock,[24] there has been several attempts to make it from lignocellulose and chitin. For the former, a one-step protocol to convert hemicellulose derived furfural and NH_3 into pyrrole was established over a Pd catalyst encapsulated within



Scheme 1. a) Representative natural products and bioactive compounds with pyrrole-2-carboxyl fragment; b) Synthetic routes of PCA in literature.

the pores of S-1 zeolite, but the catalytic system is not stable.[25] For the latter, chitin was treated with high concentration of base solution at 100 °C. The maximum yield was only 2 % when no external NH3 is added (Scheme 1b).[26] Neither lignocellulose nor chitin alone provide satisfactory synthetic routes to produce PCA.

In this study, we employed automated analysis approach to explore new routes for PCA synthesis and potential application of similar biomass-derived substrates to make a range of PCA derivatives. Aiming to deal with variable data quality reported in the literature, analyze large numbers of reactions with very different types of molecular structures and rank the identified routes against a set of pre-defined criteria, we implemented several algorithms in addition to the previously developed search and ranking of synthetic routes, including: removing hazardous molecules (intermediates, reagents, solvents), keeping/removing specific molecular fragments, automatic search of platform molecules, identifying main reactants, and 'end-to-end search'. These functions define the project-oriented search query, rather than the molecule-oriented query. With the most promising route identified, optimized, and mechanistically explored for PCA synthesis, we further created a library of valuable nitrogen-containing chemicals from PCA with particular focus on the preparation of several pyrrole-2-carboxamide/carboxylate derivatives employing 1,1′-carbonyldiimidazole (CDI) as a green coupling reagent, and several totally bio-derived N-containing heterocycles with potential use in drug discovery.

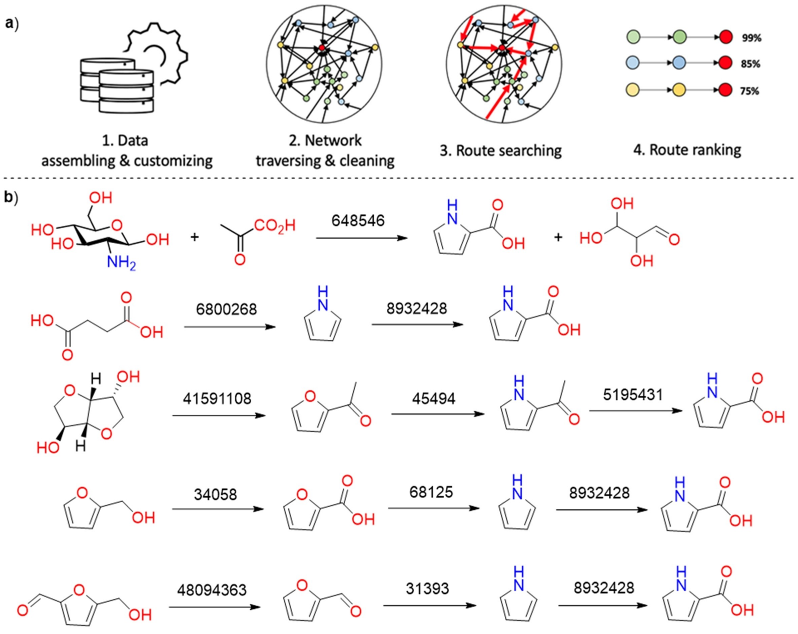
Results and Discussion

Search of synthetic routes from platform chemicals to PCA

The aim of the automated route analysis ('project') was defined as 'finding new routes to pyrrole-2-carboxylic acid (PCA) starting from cellulose and chitin bio-refineries, through datamining Reaxys®' The synthetic routes to PCA were explored by using our tools based on graphs.[27] Given a search query which includes target molecules, preferred feedstocks, preferred number of steps, and other synthetic preferences, such as avoiding chlorinated solvents, search of possible routes was performed by following four consecutive steps (see Scheme 2a):

1) assemble and customize reaction data,
2) build and clean a reaction network,
3) search synthetic routes within the network,
4) rank synthetic routes against the pre-defined criteria/indicators.

Details of the method are described in Supporting Information (SI). Synthesis planning was performed using traversal of a network in which molecules are nodes and reaction relationships are edges. Starting from one/multiple molecules, the search can be performed in a forward direction (from reactants to products) or a backward direction. A set of 357 platform chemicals was assembled through forward search from cellulose and chitin for two reaction steps, using several search heuristics, see Figure S1. Synthetic routes to PCA were searched in the backward manner for the maximum five steps. Then all possible routes connecting platform chemicals that are linked to PCA and could be obtained from cellulose and chitin were collected, filtered, and ranked by heuristics designed on the basis of molecular structures and the available reaction information.



Scheme 2. a) Workflow of identifying synthetic routes based on network analysis; b) five notable synthetic routes of PCA using platform chemicals as feedstocks identified by our automatic search tool. Number on arrows are reaction IDs in Reaxys database.

Through this effort, five routes attracted our attention, these are shown in Scheme 2b. The first path is the most promising one and will be discussed in detail later. The second path is a 2-step route starting from succinic acid that could be derived from cellulose. However, the first step in this route (Reaxys ID: 6800268, yielding pyrrol), originally reported in 1904, utilized zinc and ammonia as co-reactants that require to overcome substantial challenge to make the process catalytic. In the third path, the scaffold of PCA can be directly obtained from isosorbital (Reaxys ID: 41591108). The oxygen in the heterocyclic ring was substituted by nitrogen simply by heating 2-acetylfuran in ethanol solution of NH_3 (Reaxys ID: 45494). Nevertheless, one carbon atom was removed through oxidation using tert-butyl hydroperoxide (TBHP) as a reagent which may not be preferred due to its hazardous nature. The following two routes starting from furfuryl alcohol and 5-hydroxymethyl-2-furfuraldehyde, respectively, were not efficient because carbon atoms were initially removed and added back later in the reaction sequences.

Compared to other routes, the condensation between $GlcNH_2$ and PA was the most promising path (shown as the first one in Scheme 2b). It only requires one step, and the co-product tetrose is also a useful C_4 platform compound.[28] The reaction was originally reported in 1955 by Gottschalk. Trace amounts of PCA (0.05%) were observed when mucoprotein hydrolysates prepared from bovine submaxillary mucin was treated with an alkaline solution at 100°C. Following that, condensation reaction between glucosamine hydrochloride with sodium pyruvate in the presence of sodium carbonate was conducted, yielding about 1% of PCA.[29] By employing barium hydroxide ($Ba(OH)_2$) as a base, the PCA yield was improved to 20% based on UV/Vis spectroscopy analysis.[30] Surprisingly, no further work on this reaction has been seen in the open literature since then.

While these classical studies showed the intriguing potential of synthesizing *aza*-heterocycles from $GlcNH_2$ and PA, various issues remain. The maximum yield of PCA, even before purification, was only 20%. Second, there was no mechanistic study on the reaction pathway, thus it remains unclear how does $GlcNH_2$ and PA react to form the product. It is also unclear whether such a methodology is extendable to a broad range of substituted pyrrole-2-carboxylic acids. Finally, since PCA is not purified, its potential as the starting material to make molecules bearing more complex structure is not explored.

Optimization of PCA synthesis from D-glucosamine hydrochloric acid ($GlcNH_2 \cdot HCl$) and PA

We investigated the influence of bases on the reaction between $GlcNH_2 \cdot HCl$ and PA to form PCA. Lithium hydroxide (LiOH) gave a higher yield (19.4%) than sodium hydroxide (NaOH) (8.4%), potassium hydroxide (KOH) (16.7%), barium hydroxide ($Ba(OH)_2$) (13.4%) and potassium carbonate (K_2CO_3) (2.3%). Since PA undergoes both self-condensation and degradation under strongly basic conditions, the yield of PCA increased from 19.4% to 29.7% when the ratio between $GlcNH_2 \cdot HCl$ and PA increased from 1:3 to 1:6 (Table 1, entry 4, 7 and 8). The applicability of alternative solid bases including MgO, CaO and hydrotalcite was also examined. In these cases, sodium hydroxide was first used to convert PA to pyruvate salt in situ before adding the solid catalyst. No product was obtained when MgO was used but the PCA yield reached 7% in the presence of hydrotalcite, and 26.5% in the presence of CaO that is a stronger base, suggesting the critical role of catalyst basicity (Table 1, entry 9).

The reaction was then conducted at different temperatures while keeping other parameters unchanged. There was no product observed by thin layer chromatography at 60°C. The PCA yield significantly increased with temperature increased from 70°C (7.8%) to 80°C (27.2%), and further increased when temperature was increased to 120°C, leading to the highest yield of 34.9% (Figure 1a). When followed as a function of time at various temperatures, similar trends were observed for reactions at 80 and 100°C; the formation of PCA increases steadily from 30 to 240 min, with the highest isolated yield (38.8%) obtained at 100°C over 240 min. In contrast, reaction at 120°C reached its peak after one hour (37.3%) and declined gradually afterwards (Figure 1b).

Based on these results, we hypothesized that the fed batch operation would further enhance the yield, considering the product is not stable. In an initial trial, a mixture of PA, $GlcNH_2 \cdot HCl$, and LiOH in H_2O were dropwise feed to preheated water (100°C) over a period of 210 min, followed by another 30-min reaction, but only 25.1% yield of PCA was obtained, highlighting a high concentration alkaline condition is critical for this cyclization reaction. Subsequently, we continuously fed PA and $GlcNH_2 \cdot HCl$ to a pre-added basic aqueous solution, resulting in 49.6% yield of the target product. Finally, the effect of varying feeding times during the reaction was investigated (Figure 2), and the highest isolated yield (50.8%) of PCA was obtained.

Table 1. Effect of bases and substrates ratio on PCA formation.[a]

Entry	$GlcNH_2 \cdot HCl$:PA	Base	Isolated yield (%)
1	1:3	–	–
2	1:3	NaOH	8.4
3	1:3	KOH	16.7
4	1:3	LiOH	19.4
5	1:3	$Ba(OH)_2$	13.4
6	1:3	K_2CO_3	2.3
7	1:5	LiOH	27.7
8	1:6	LiOH	29.7
9	1:6	CaO[b]	26.5
10	1:6	MgO[b]	–
11	1:6	hydrotalcite[b]	7.0

[a] Reaction conditions: H_2O, 100°C, 2 h. PCA yield = mol of PCA per mol of $GlcNH_2 \cdot HCl \times 100\%$; the amount of PCA is weighed after purification and drying *in vacuo* at room temperature overnight. [b] NaOH was first used to convert PA to sodium pyruvate.

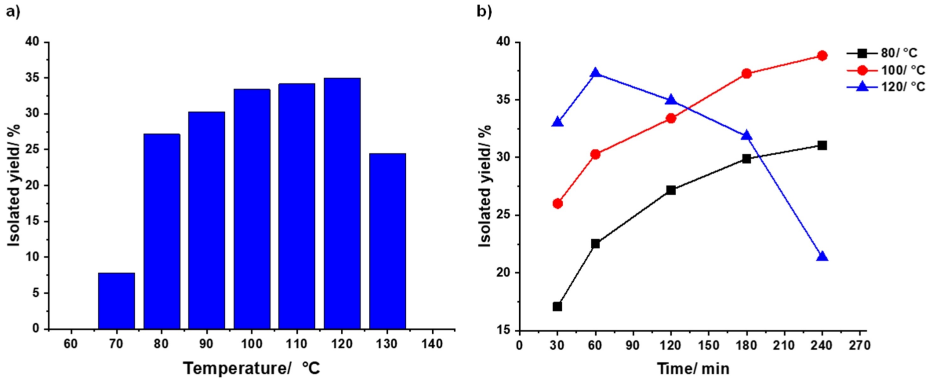


Figure 1. a) Screening different temperatures. Reaction conditions: $GlcNH_2 \cdot HCl$:PA (1 : 6), LiOH (12 equiv.), H_2O, 2 h; b) Reaction over times at different temperatures. Reaction conditions: $GlcNH_2 \cdot HCl$:PA (1 : 6), LiOH (12 equiv.), H_2O.

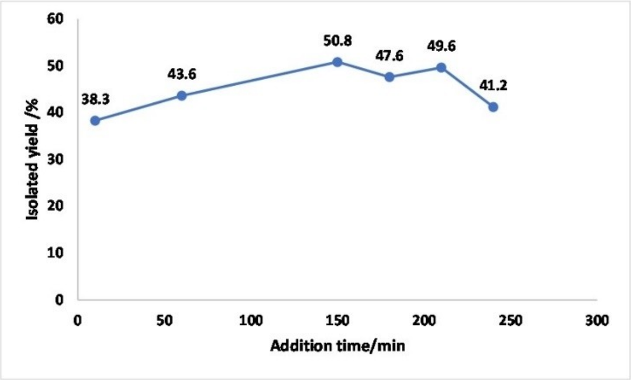


Figure 2. Screening of different addition times. Reaction conditions: $GlcNH_2 \cdot HCl$ (1 equiv.) and PA (6 equiv.) in H_2O was added dropwise to a mixture of LiOH (12 equiv.) in H_2O at 100 °C and stirred for 240 min.

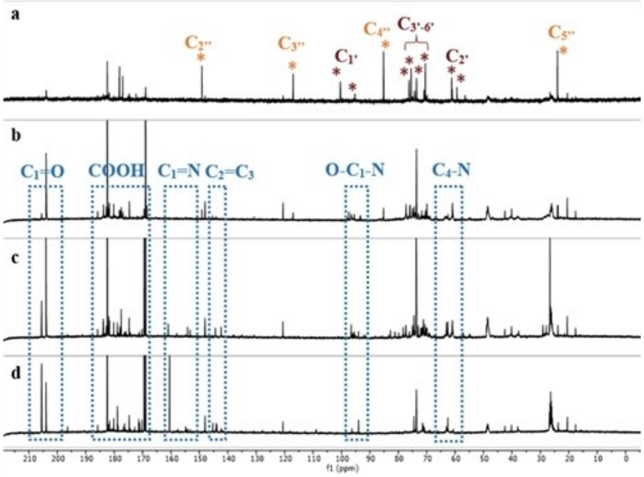
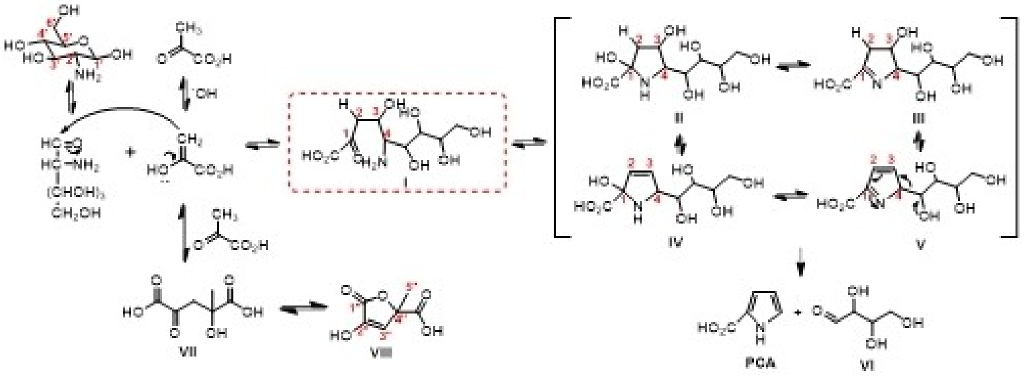


Figure 3. ^{13}C NMR spectra of reaction mixture at different temperatures. a) 30 °C; b) 60 °C, c) 80 °C; d) 120 °C. Reaction conditions: $GlcNH_2$:PA (1 : 6), NaOH (12 equiv.), D_2O, 30 min. (Labels based on the proposed mechanism shown in Scheme 3).

Reaction pathway y illustration

We conducted controlled experiments and ^{13}C NMR spectroscopic analysis to identify intermediates and elucidate reaction pathway (Scheme 3). The ^{13}C NMR spectra were recorded for the reaction mixture conducted at 30 °C, 60 °C, 80 °C and 120 °C after 30 min (Figure 3). One peak at 205 ppm appears in the spectrum of reaction mixtures at 60 °C and above, which is absent in the spectrum for the sample after exposure at room temperature. We assign it to the α-ketone group in compound **I** formed from the aldol condensation between $GlcNH_2$ and PA. To verify this, control reactions with only $GlcNH_2$ or PA as



Scheme 3. Proposed mechanism for PCA formation from $GlcNH_2$ and pyruvic acid.

substrate were carried out under the same conditions. This peak did not appear in neither case. Further, we replaced $GlcNH_2$ with NAG and glucose to react with pyruvic acid. For these experiments, the same peak also appeared (Figure S2). These control experiments provide strong evidence that the new peak is a result of aldol adduct formation between PA and the sugar unit. The reaction at room temperature did not show any sign for the formation of aldol adduct **I**; instead, the peaks of $GlcNH_2$ in the ring forms (α- and β-anomers) were still observed in the spectrum (Figure 3a, brown stars). PA followed self-condensation route to form compound **VII**, which underwent lactonization to form lactone **VIII** (Figure 3a, orange stars).[31]

We also detected salient peaks in the region between 93 and 97 ppm. These peaks are ascribed to the carbon in the hemiaminal O–C1–N bond belonging to intermediates **II**, formed by nucleophilic attack of the amino group to the carbonyl group in **I**. This hemiaminal O–C1–N bond could also belong to intermediate **IV**, which is a dehydrated product of **II**. It is noted that there were no similar peaks in these regions in the case of N-acetyl glucosamine and glucose as substrates, indicating the free amino group is crucial for the formation of imine bond (Figure S1). The formation of intermediate **III** and **V**, which resulted from dehydration of **II** and **IV**, respectively, was supported by the appearance of peaks at 150–160 ppm belonging to imine C=N bonds. The formation of intermediate **IV** and **V** could also be based on the C2=C3 double bond's expected peaks at 142–146 ppm shown in the 80 °C and 120 °C spectra. There are no peaks in these regions when the reaction was conducted at 30 °C and 60 °C, highlighting the requirement of high temperature for dehydration, in agreement with our temperature effect study showing no PCA formation at 60 °C. The last step is the cleavage of the side chain which could occur by *retro*-aldol type to form PCA and by-product aldehyde **VI**. However, attempts to detect aldehyde **VI** by ESI-MS was unsuccessful, suggesting the instability of **VI** under reaction conditions.

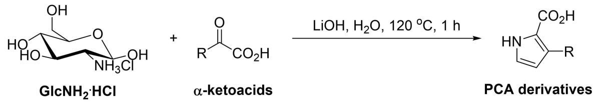
Synthesis of PCA derivatives from $GlcNH_2$ and biomass-derived α-ketoacids

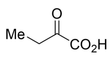
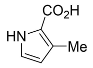
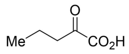
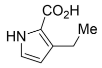
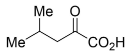
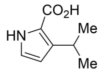
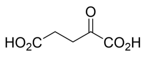
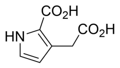
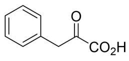
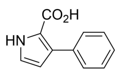
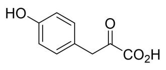
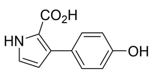
With the reaction conditions optimized, several α-keto acids were applied to expand the scope of the reaction to make 3-substituted pyrrole-2-carboxylic acids, whose synthesis usually requires complex strategies from petroleum-based substrates.[32] We focused on biomass-derived α-keto acids to prepare fully bio-based chemicals. Expectedly, the reaction of α-ketobutyric acid **1** with $GlcNH_2$ results in the formation of 3-methylpyrrole-2-carboxylic acid **7** with a reasonable yield of 42%. On the other hand, the reaction of $GlcNH_2$ with bio-derived 4-methyl-2-oxovaleric acid **3** did not lead to isopropyl derivative **9**, likely due to the steric hindrance from the isopropyl group in **3**. To verify this, we conducted the reaction between $GlcNH_2$ and 2-oxovaleric acid **2** bearing smaller steric hindrance, and indeed the desired product 3-ethylpyrrole-2-carboxylic acid **8** was obtained. Dicarboxylic **10** was easily prepared from α-ketoglutaric acid **4** and $GlcNH_2$ in 29% yield. Bio-derived phenyl derivatives including phenylpyruvic acid **5** and 4-hydroxyphenylpyruvic acid **6** were also reacted with $GlcNH_2$ giving the bicyclic derivatives **11** and **12**, respectively, which could potentially be used as substrates for the synthesis of Kendine 91 (structure shown in Scheme 1a) and its derivatives for drug discovery. The successful synthesis of these derivatives offers a direct pathway to prepare 2,3-substituted pyrrole derivatives from petroleum-free and Haber-Bosh-independent substrates at reasonable yields (25-48%) (Table 2).

Synthesis of N-containing compounds from PCA

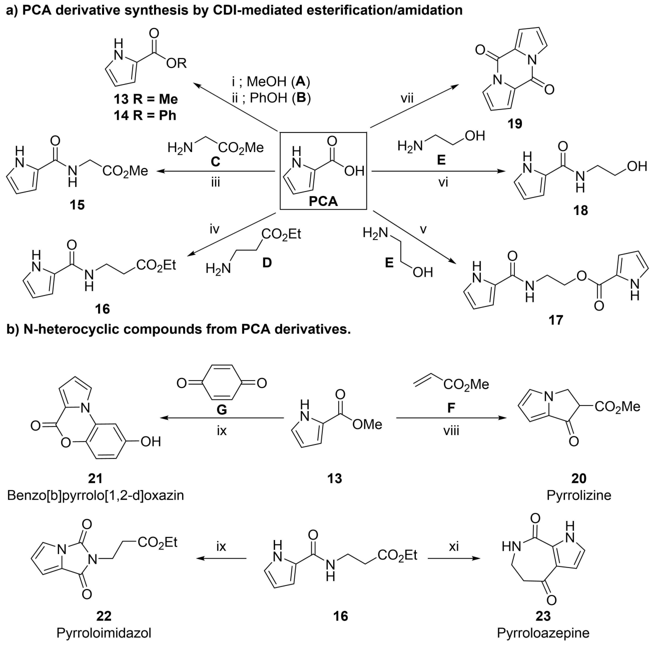
Our next target was to establish a library of nitrogen-containing compounds from the bio-based PCA. As PCA-derived esters and amides are common moieties present in the pyrrole alkaloids, we first synthesized a range of ester and amide derivatives (Scheme 4a). For example, by using biomass-derived alcohols (**A** and **B**) and amines (**C**, **D** and **E**) as reagents and 1,1′-carbonyldiimidazole (CDI) as a green coupling reagent in acetonitrilethe, PCA could be transformed into full bio-based esters (and) and amides (**15** to **19**).[33] In addition, the synthesis of several drug-like fused heterocycles from PCA and other bio-derived molecules was investigated (Scheme 4b). Methyl ester **13** underwent aza Michael addition with bio-derived methyl acrylate **F** and subsequent Dieckmann condensation in the

Table 2. Synthesis of 3-substituted pyrrole-2-carboxylic acid derivatives from $GlcNH_2 \cdot HCl$ and biomass-derived α-keto acid[a].



α-ketoacids	PCA derivatives	Isolated yield (%)
 **1**	 **7**	42
 **2**	 **8**	25
 **3**	 **9**	--
 **4**	 **10**	29
 **5**	 **11**	48
 **6**	 **12**	37

[a] Reaction conditions: $GlcNH_2 \cdot HCl$ (1.0 equiv.), PA (6.0 equiv.), LiOH (12.0 equiv.), H_2O, 0 °C~120 °C, 1.0 h.



Scheme 4. Conversion of PCA into valuable chemicals containing nitrogen. Reagents and conditions (i to vi): CDI, Et_3N, MeCN, 100 °C, 1 h; i) then **A**, 100 °C, 2 h, 76 %; ii) then **B**, 100 °C, 3 h, 92 %; iii) then **C**, 100 °C, 6 h, 58 %; iv) then **D**, 100 °C, 6 h, 91 %; v) then **E** (0.5 equiv.), 100 °C, 12 h, 38 %; vi) then **E** (1.1 equiv.), 100 °C, 6 h, 75 %; vii) then 140 °C, 8 h, 45 %; reaction conditions: viii) a. **F**, K_2CO_3, MeCN, 100 °C, 1 h, b. NaH, toluene, 130 °C, 2 h, 56 %; ix) G, K_2CO_3, MeCN, 110 °C, 3 64 %; x) CDI, Et_3N, toluene, 130 °C, 4 h, 92 %; xi) triflic acid, 100 °C, 1 h, 90 %.

presence of sodium hydride to obtain pyrrolizine **20**. Tandem *aza*-Michael addition/lactonization between methyl ester **13** and *p*-benzoquinone **G** occurred to give the tricyclic **21** under simple conditions using only K_2CO_3 as a base. Next, amide **16** was easily reacted with CDI to form pyrroloimidazole **22** with high yield (92 %). Finally, amide **16** was subjected to intramolecular cyclisation mediated by triflic acid to give pyrroloazepine **23**, which is a main core structure of the biologically active natural aldisine alkaloids.[34]

Conclusions

In summary, we have used automated synthesis planning based on reaction networks evaluation to identify a key reaction step to establish a synthesis route to valuable organonitrogen chemicals that is based on glucosamine and pyruvic acid, sourced from two types of biorefineries – cellulose and chitin. The key intermediate, PCA was obtained with the isolated yield close to 40 %, while the ^{13}C NMR study identified aldol adducts and other intermediates which allowed to construct a plausible reaction pathway. The scope of this reaction was readily expanded to other biomass-derived a-ketoacids, resulting in the synthesis of various PCA derivatives substituted at 3-position. Moreover, we demonstrate that PCA can become a platform molecule to make important derivatives such as alcohols, amines and drug-like pyrrolizine, pyrroloimidazole, pyrroloazepine and benzopyrrolooxazine derivatives.

The work shown here highlights the effectiveness of automated synthesis planning based on existing databases to assist in design of synthesis routes that require non-trivial combination of raw materials. This study is one of few that illustrates the full workflow of CASP followed by experimental verification and reaction scope expansion, pathing the way to more common use of CASP by experimental chemists.

Experimental Section

To a solution of PA in water at 0 °C was slowly added solution of base in water. After that, the mixture was heated at a specific time and was drop-wised solution of D-glucosamine.HCl in water within 10 min. The mixture was stirred at a specific temperature for a specific time. After completion, the reaction was acidified with HCl 6 M upon pH 3–4 (paper pH indicator). The mixture was extracted with diethyl ether (3 times). The combined extract was dried over Na2SO4 and concentrated in vacuo. The residue was purified by column chromatography eluting with ethyl acetate: n-hexane (1 : 1) to obtain the desired product.

Supporting Information

Detailed methodology on automatic searching of synthetic routes from platform, and NMR spectra of various products. This material is available free of charge via the Internet at http://pubs.acs.org.

Acknowledgements

We are grateful to the financial support provided by the National Research Foundation (NRF) of Singapore under its Campus for Research Excellence and Technological Enterprise (CREATE) program (CARES), as well as the C4T project under CARES (WBS R-279-000-604-592). We acknowledge collaboration with RELX Intellectual Properties SA and their technical support, which enabled us to mine Reaxys®. Reaxys® data were made accessible to our research project via the Elsevier R&D Collaboration Network.

Conflict of Interests

The authors declare no conflict of interest.

Data Availability Statement

The data that support the findings of this study are available from the corresponding author upon reasonable request.

Keywords: biorefinery · computer assisted synthesis planning · green chemistry · reaction networks

[1] A. Corma, S. Iborra, A. Velty, *Chem. Rev.* **2007**, *107*, 2411.

[2] a) L. T. Mika, E. Csefalvay, A. Nemeth, *Chem. Rev.* **2018**, *118*, 505; b) Z. Zhou, D. Liu, X. Zhao, *Renewable Sustainable Energy Rev.* **2021**, *146*, 111169; c) H. Xu, Z. Wang, J. Huang, Y. Jiang, *Energy Fuels* **2021**, *35*, 8602; d) F. Pattnaik, S. Tripathi, B. R. Patra, S. Nanda, V. Kumar, A. K. Dalai, S. Naik, *Environ. Chem. Lett.* **2021**, *19*, 4119; e) V. P. Santos, N. S. S. Marques, P. Maia, M. A. B. Lima, L. O. Franco, G. M. Campos-Takaki, *Int. J. Mol. Sci.* **2020**, *21*, 4290; f) E. Nzediegwu, M.-J. Dumont, *Waste Biomass Valorization* **2020**, *12*, 2825; g) J. Dai, F. Li, X. Fu, *ChemSusChem* **2020**, *13*, 6498.

[3] a) A. Galadima, O. Muraza, *Fuel* **2019**, *258*, 115851; b) K. Kohli, R. Prajapati, B. Sharma, *Energies* **2019**, *12*, 1; c) C. H. L. Tempelman, R. Oozeerally, V. Degirmenci, *Catalysts* **2021**, *11*, 861.

[4] a) Z. Jiang, D. Hu, Z. Zhao, Z. Yi, Z. Chen, K. Yan, *Processes* **2021**, *9*, 1234; b) N. A. S. Ramli, N. A. S. Amin, *BioEnergy Res.* **2020**, *13*, 693; c) C. Xu, E. Paone, D. Rodriguez-Padron, R. Luque, F. Mauriello, *Chem. Soc. Rev.* **2020**, *49*, 4273.

[5] a) W. Fan, C. Verrier, Y. Queneau, F. Popowycz, *Curr. Org. Synth.* **2019**, *16*, 583; b) L. Hu, W. Zhen, Y. Jiang, X. Wang, A. He, J. Song, J. Xu, S. Zhou, Y. Zhao, J. Xu, *Renewable Sustainable Energy Rev.* **2020**, *134*, 110317.

[6] a) Z. Luo, S. Yu, W. Zeng, J. Zhou, *Biotechnol. Adv.* **2021**, *47*, 107706; b) C. Yin, X. Li, Z. Chen, W. Zou, Y. Peng, S. Wei, C. Tang, L. Dong, *New J. Chem.* **2020**, *44*, 5884.

[7] a) X. Chen, H. Yang, N. Yan, *Chem. Eur. J.* **2016**, *22*, 13402; b) M. J. Hülsey, H. Yang, N. Yan, *ACS Sustainable Chem. Eng.* **2018**, *6*, 5694; c) T. T. Pham, A. C. Lindsay, X. Chen, G. Gözaydin, N. Yan, J. Sperry, *Chem. Pharm.* **2019**, *13*, 100143; d) M. Nikahd, J. Mikusek, L. J. Yu, M. L. Coote, M. G. Banwell, C. Ma, M. G. Gardiner, *J. Org. Chem.* **2020**, *85*, 4583; e) T. T. Pham, X. Chen, T. Söhnel, N. Yan, J. Sperry, *Green Chem.* **2020**, *22*, 1978.

[8] a) C. W. Coley, W. H. Green, K. F. Jensen, *Acc. Chem. Res.* **2018**, *51*, 1281; b) C. W. Coley, W. H. Green, K. F. Jensen, *J. Chem. Inf. Model.* **2019**, *59*, 2529; c) Y. Mo, Y. Guan, P. Verma, J. Guo, M. E. Fortunato, Z. Lu, C. W. Coley, K. F. Jensen, *Chem. Sci.* **2020**, *12*, 1469; d) https://askcos.mit.edu/.

[9] a) M. H. S. Segler, M. Preuss, M. P. Waller, *Nature* **2018**, *555*, 604; b) http://www.elsevier.com/solutions/reaxys/features-and-capabilities/retrosynthesis.

[10] a) P. Schwaller, R. Petraglia, V. Zullo, V. H. Nair, R. A. Haeuselmann, R. Pisoni, C. Bekas, A. Iuliano, T. Laino, *Chem. Sci.* **2020**, *11*, 3316; b) P. Schwaller, A. C. Vaucher, T. Laino, J.-L. Reymond, *Mach. Learn.: Sci. Technol.* **2021**, *2*, 015016.

[11] a) A. A. Lee, Q. Yang, V. Sresht, P. Bolgar, X. Hou, J. L. Klug-McLeod, C. R. Butler, *Chem. Commun. (Camb.)* **2019**, *55*, 12152; b) https://postera.ai/manifold/.

[12] a) S. Szymkuc, E. P. Gajewska, T. Klucznik, K. Molga, P. Dittwald, M. Startek, M. Bajczyk, B. A. Grzybowski, *Angew. Chem. Int. Ed.* **2016**, *55*, 5904; *Angew. Chem.* **2016**, *128*, 6004; b) T. Klucznik, B. Mikulak-Klucznik, M. P. McCormack, H. Lima, S. Szymkuć, M. Bhowmick, K. Molga, Y. Zhou, L. Rickershauser, E. P. Gajewska, A. Toutchkine, P. Dittwald, M. P. Startek, G. J. Kirkovits, R. Roszak, A. Adamski, B. Sieredzińska, M. Mrksich, S. L. J. Trice, B. A. Grzybowski, *Chem* **2018**, *4*, 522; c) B. Mikulak-Klucznik, P. Golebiowska, A. A. Bayly, O. Popik, T. Klucznik, S. Szymkuc, E. P. Gajewska, P. Dittwald, O. Staszewska-Krajewska, W. Beker, T. Badowski, K. A. Scheidt, K. Molga, J. Mlynarski, M. Mrksich, B. A. Grzybowski, *Nature* **2020**, *588*, 83; d) T. Badowski, E. P. Gajewska, K. Molga, B. A. Grzybowski, *Angew. Chem. Int. Ed.* **2020**, *59*, 725; *Angew. Chem.* **2020**,*132*,735.

[13] a) M. Fialkowski, K. J. M. Bishop, V. A. Chubukov, C. J. Campbell, B. A. Grzybowski, *Angew. Chem. Int. Ed.* **2005**, *44*, 7263; *Angew. Chem.* **2005**, *117*, 7429; b) B. A. Grzybowski, K. J. M. Bishop, B. Kowalczyk, C. E. Wilmer, *Nat. Chem.* **2009**, *1*, 31; c) S. Szymkuc', E. P. Gajewska, T. Klucznik, K. Molga, P. Dittwald, M. Startek, M. Bajczyk, B. A. Grzybowski, *Angew. Chem. Int. Ed.* **2016**, *55*, 5904; *Angew. Chem.* **2016**, *128*, 6004; d) B. Mikulak-Klucznik, P. Gołębiowska, A. A. Bayly, O. Popik, T. Klucznik, S. Szymkuć, E. P. Gajewska, P. Dittwald, O. Staszewska-Krajewska, W. Beker, T. Badowski, K. A. Scheidt, K. Molga, J. Mlynarski, M. Mrksich, B. A. Grzybowski, *Nature* **2020**, *588*, 83; e) M. Kowalik, C. M. Gothard, A. M. Drews, N. A. Gothard, A. Weckiewicz, P. E. Fuller, B. A. Grzybowski, K. J. M. Bishop, *Angew. Chem. Int. Ed.* **2012**, *51*, 7928; *Angew. Chem.* **2012**, *124*, 8052.

[14] P. M. Jacob, P. Yamin, C. Perez-Storey, M. Hopgood, A. A. Lapkin, *Green Chem.* **2017**, *19*, 140.

[15] J. M. Weber, P. Lió, A. A. Lapkin, *React. Chem. Eng.* **2019**, *4*, 1969.

[16] J. Dong, M. Zhao, Y. Liu, Y. Su, X. Zeng, *Briefings Bioinf.* **2022**, *23*, bbab391.

[17] B. A. Grzybowski, T. Badowski, K. Molga, S. Szymkuć, *WIREs Comput. Mol. Sci.* **2023**, *13*, e1630.

[18] J. M. Weber, Z. Guo, C. Zhang, A. M. Schweidtmann, A. A. Lapkin, *Chem. Soc. Rev.* **2021**, *50*, 12013.

[19] J. M. Weber, Z. Guo, A. A. Lapkin, *ACS Eng. Au.* **2022**, *2*, 333.

[20] A. Arun, Z. Guo, S. Sung, A. A. Lapkin, *Chemistry–Methods* **2023**, *3*, e202200062.

[21] a) B. Forte, B. Malgesini, C. Piutti, F. Quartieri, A. Scolaro, G. Papeo, *Mar. Drugs* **2009**, *7*, 705; b) R. Rane, C. Shah, R. Karpoormath, *Curr. Top. Med. Chem.* **2014**, *14*, 253; c) F. Scala, E. Fattorusso, M. Menna, O. Taglialatela-Scafati, M. Tierney, M. Kaiser, D. Tasdemir, *Mar. Drugs* **2010**, *8*, 2162; d) K. Seipp, L. Geske, T. Opatz, *Mar. Drugs* **2021**, *19*, 514; e) H. Zhang, M. Dong, J. Chen, H. Wang, K. Tenney, P. Crews, Agelas, *Mar. Drugs* **2017**, *15*, 351.

[22] a) A. E. Cotman, T. Guerin, I. Kovacevic, D. Benedetto Tiz, M. Durcik, F. Fulgheri, S. Mozina, D. Secci, M. Sterle, J. Ilas, A. Zega, N. Zidar, L. P. Masic, T. Tomasic, F. R. Leroux, G. Hanquet, D. Kikelj, *ACS Omega* **2021**, *6*, 9723; b) G. S. Basarab, P. J. Hill, C. E. Garner, K. Hull, O. Green, B. A. Sherer, P. B. Dangel, J. I. Manchester, S. Bist, S. Hauck, F. Zhou, M. Uria-Nickelsen, R. Illingworth, R. Alm, M. Rooney, A. E. Eakin, *J. Med. Chem.* **2014**, *57*, 6060.

[23] A. Zubia, S. Ropero, D. Otaegui, E. Ballestar, M. F. Fraga, M. Boix-Chornet, M. Berdasco, A. Martinez, L. Coll-Mulet, J. Gil, F. P. Cossio, M. Esteller, *Oncogene* **2009**, *28*, 1477.

[24] D. Tzankova, S. Vladimirova, L. Peikova, M. Georgieva, *J. Chem. Technol. Metall.* **2018**, *53*, 451.

[25] S. Song, V. F. K. Yuen, L. Di, Q. Sun, K. Zhou, N. Yan, *Angew. Chem. Int. Ed.* **2020**, *59*, 19846; *Angew. Chem.* **2020**,*132*, 20018.

[26] X. Gao, X. Chen, J. Zhang, W. Guo, F. Jin, N. Yan, *ACS Sustainable Chem. Eng.* **2016**, *4*, 3912.

[27] a) J. M. Weber, Z. Guo, C. Zhang, A. M. Schweidtmann, A. A. Lapkin, *Chem. Soc. Rev.* **2021**, *50*, 12013; b) J. M. Weber, Z. Guo, A. A. Lapkin, *ACS Engineering Au* **2022**, *2*, 333.

[28] a) M. Dusselier, P. V. Wouwe, F. de Clippel, J. Dijkmans, D. W. Gammon, B. F. Sels, *ChemCatChem* **2013**, *5*, 569; b) A. Mollar-Cuni, J. P. Byrne, M. P. Borja, C. Vicent Barrera, M. Albrecht, J. A. Mata, *ChemCatChem* **2020**, *12*, 3746.

[29] a) A. Gottschalk, *Nature* **1955**, *176*, 881; b) G. Cooper, C. Reed, D. Nguyen, M. Carter, Y. Wang, *Proc. Natl. Acad. Sci. USA* **2011**, *108*, 14015.

[30] A. Gottschalk, *Arch. Biochem. Biophys.* **1957**, *59*, 37.

[31] A. C. Rios, P. P. Bera, J. A. Moreno, G. Cooper, *ACS Omega* **2020**, *5*, 15063.

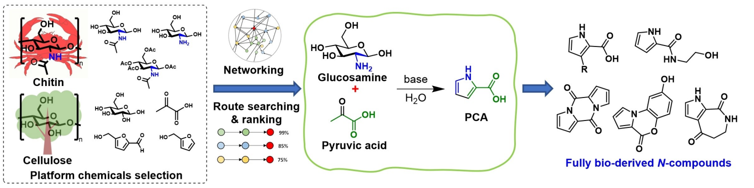
[32] V. Desplat, S. Moreau, A. Gay, S. B. Fabre, D. Thiolat, S. Massip, G. Macky, F. Godde, D. Mossalayi, C. Jarry, J. Guillon, *J. Enzyme Inhib. Med. Chem.* **2010**, *25*, 204.

[33] J. P. Adams, C. M. Alder, I. Andrews, A. M. Bullion, M. Campbell-Crawford, M. G. Darcy, J. D. Hayler, K. R. Henderson, C. A. Oare, I. Pendrak, A. M. Redman, L. E. Shuster, H. F. Sneddon, M. D. Walker, *Green Chem.* **2013**, *15*, 1542.

[34] a) R. Akunuri, M. Vadakattu, S. Bujji, V. Veerareddy, Y. V. Madhavi, S. Nanduri, *Eur. J. Med. Chem.* **2021**, *220*, 113445; b) D. Tasdemir, R. Mallon, M. Greenstein, L. R. Feldberg, S. C. Kim, K. Collins, D. Wojciechowicz, G. C. Mangalindan, G. P. Concepción, M. K. Harper, C. M. Ireland, *J. Med. Chem.* **2002**, *45*, 529.

Manuscript received: August 10, 2023
Revised manuscript received: October 2, 2023
Accepted manuscript online: October 4, 2023
Version of record online: ■■, ■■

RESEARCH ARTICLE



In this work, a protocol was developed to identify potential transformation pathways that allow for the use of feedstocks from different origins. After an extensive analysis of chemical reactions recorded in extensive databases, several promising synthetic routes were shortlisted, with the reaction of D-glucosamine and pyruvic acid being the most interesting to make pyrrole-2-carboxylic acid (PCA). The optimized synthetic conditions resulted in 50 % yield of PCA, with insights gained from temperature variant NMR studies.

Dr. T. T. Pham, Dr. Z. Guo, Dr. B. Li, Prof. A. A. Lapkin, Prof. N. Yan**

1 – 9

Synthesis of Pyrrole-2-Carboxylic Acid from Cellulose- and Chitin-Based Feedstocks Discovered by the Automated Route Search