JOURNAL OF
CHEMICAL INFORMATION
AND MODELING

Article

pubs.acs.org/jcim

Beyond Terrestrial Biology: Charting the Chemical Universe of α-Amino Acid Structures

Markus Meringer,[†] H. James Cleaves II,*[,‡,§,⊥,∥] and Stephen J. Freeland[○]

[†]German Aerospace Center (DLR), Earth Observation Center (EOC), Münchner Straße 20, D-82234 Oberpfaffenhofen−Wessling, Germany

[‡]Earth-Life Science Institute, Tokyo Institute of Technology, 2-12-1 Ookayama, Meguro-ku, Tokyo 152-8550, Japan

[§]Institute for Advanced Study, 1 Einstein Drive, Princeton, New Jersey 08540, United States

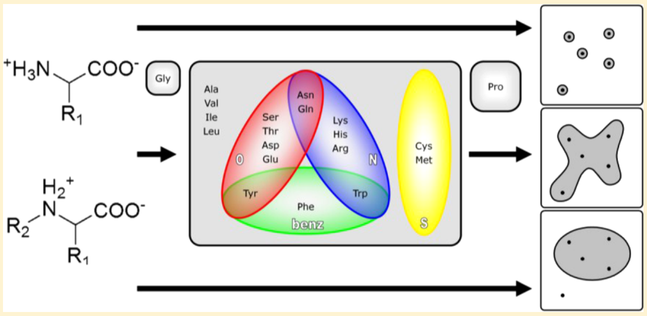
[⊥]Blue Marble Space Institute of Science, 2800 Woodley Road NW, no. 544, Washington, D.C. 20016, United States

[∥]Center for Chemical Evolution, Georgia Institute of Technology, Atlanta, Georgia 30332, United States

[○]NASA Astrobiology Institute, University of Hawaii, 2680 Woodlawn Drive, Honolulu, Hawaii 96822-1839, United States

S *Supporting Information*

ABSTRACT: α-Amino acids are fundamental to biochemistry as the monomeric building blocks with which cells construct proteins according to genetic instructions. However, the 20 amino acids of the standard genetic code represent a tiny fraction of the number of α-amino acid chemical structures that could plausibly play such a role, both from the perspective of natural processes by which life emerged and evolved, and from the perspective of human-engineered genetically coded proteins. Until now, efforts to describe the structures comprising this broader set, or even estimate their number, have been hampered by the complex combinatorial properties of organic molecules. Here, we use computer software based on graph theory and constructive combinatorics in order to conduct an efficient and exhaustive search of the chemical structures implied by two careful and precise definitions of the α-amino acids relevant to coded biological proteins. Our results include two virtual libraries of α-amino acid structures corresponding to these different approaches, comprising 121 044 and 3 846 structures, respectively, and suggest a simple approach to exploring much larger, as yet uncomputed, libraries of interest.



INTRODUCTION

In one of the earliest and most significant milestones of biological evolution,[1,2] life established a set of 20 different α-amino acids with which to construct genetically encoded proteins. α-Amino acids are relatively simple organic molecules in which a carboxyl functional group is connected to an amino functional group by a single "α" carbon. This α-carbon may also be bonded to one or more side-chains of variable composition (Figure 1).

Biological systems construct proteins by joining individual amino acids into linear polymers via peptide linkages, covalent bonds that connect the carboxyl group of one amino acid to the amino group of another. A foundational insight of molecular biology is that the precise sequence by which amino acids are linked together is often sufficient to determine the 3-dimensional shape into which the resulting polypeptide ultimately folds.[3] A specific linear sequence of amino acids thus implies a particular suite of physicochemical characteristics for the resulting protein that define its activity, including its interaction with other molecules. As such, genetic material needs only specify various linear sequences of amino acids in order to generate all the proteins necessary for a self-replicating metabolism. Viewed in this way, it is remarkable that just 20 different types of amino acid have proven sufficient to build all the genetically encoded proteins of most organisms for more than 3 billion years, especially given our emerging awareness of the evironmental extremes to which various branches of life have adapted.[4] Even more remarkably, α-amino acids form a direct chemical connection to the nonliving cosmos in that they have also been found in meteorites,[5] identified in returned cometary samples,[6] and synthesized in simulations of our planet's prebiotic chemistry.[7−9] Glycine, the simplest of the genetically encoded amino acids, may even be present in the interstellar medium.[10−12]

However, nonbiological processes can and often do produce far more than 20 amino acids, including α-amino acids beyond those found in the genetic code, as well as β-, γ-, and δ-amino acids and others (see Figure 1). For example, 75 to 100 different amino acids have been detected in the Murchison meteorite to date,[13] and improvements in analytical sensitivity continue to reveal a far greater diversity of molecular structure

Received: April 6, 2013
Published: October 23, 2013

dx.doi.org/10.1021/ci400209n | *J. Chem. Inf. Model.* 2013, 53, 2851−2862

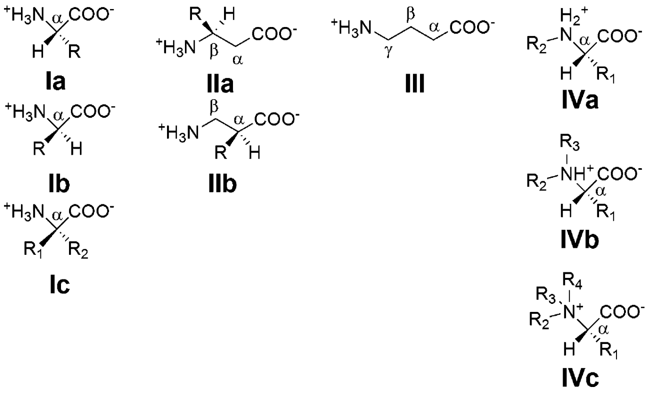


Figure 1. Generic structural types of amino acids, shown as their zwitterions with respect to their core α-amino acid motif, using standard notation that R is a side chain of variable structure and composition: (Ia) and (Ib) are the L- and D-stereoisomers of a simple α-amino acid; (Ic) is an α,α-dialkyl amino acid with two alkyl side chains bound to the α carbon; (IIa) and (IIb) show two of the many variations that become possible when an extra carbon atom is inserted between amino and carboxyl functional groups so as to form a β-amino acid; (III) is a γ-amino acid; (IVa), (IVb), and (IVc) illustrate secondary, tertiary, and quaternary amines, respectively. All genetically encoded amino acids are of type (Ia) with the exception of proline, which is of type (IVa).

than was previously suspected in both meteoritic samples and prebiotic simulations.[8,14−17] Despite this molecular diversity, the products of abiotic chemistry can account for only around half of the 20 genetically encoded amino acids,[18,19] the remainder seem to have emerged as biosynthetic modifications of simpler amino acids or other metabolites:[20] “inventions” of biological evolution that were subsequently incorporated into the genetic code. Indeed the amino acids selenocysteine and pyrrolysine, which fit this description, are currently entering the genetic code as the 21st and 22nd coded amino acid within some lineages.[21] In this context, it is noteworthy that diverse biological systems use far more types of amino acids than the 20 into which genes are decoded. These additional amino acids fall into various categories, including secondary metabolites, post-translational modifications, amino acids used in nonribosomal peptide synthesis, and intermediates of the metabolic pathways by which the standard 20 are synthesized and degraded (see Figure 2). The total number of amino acids occurring in biological systems is unknown; however, estimates range into the hundreds or thousands, with the majority found in plants and microbes (e.g., see refs 22−24).

Since abiotic synthesis and metabolism can each produce many amino acids besides those found in the genetic code,[22,24] it seems that biological evolution selected 20 for use in genetic coding from a much larger pool of possible chemical structures.[25] While the total number of possible α-amino acids is technically infinite, limiting descriptors (such as a molecular weight cap) can be used to define finite sets of α-amino acid structures. The fact that such sets have been neither defined nor have their structures been generated reflects the complexity introduced by combinatorics. While simple algorithms have been used to calculate the total number of possible alkyl amino acids,[26] the incorporation of heteroatoms (i.e., atoms other than carbon or hydrogen) vastly increases the potential for molecular diversity and the corresponding challenges for exploration.

There are at least three reasons to explore and better understand this universe of possible α-amino acid structures. One comes from astrobiology, the interdisciplinary search to understand how life emerges in the cosmos.[2] An important challenge here is to understand whether physical or chemical principles predict which 20 α-amino acids would be selected by evolution from the near-infinite number of structural possibilities. Are other possible combinations better in some obvious functional respects, such as in the coverage of physical properties which might be useful in protein folding or catalysis?[27] If so, then the outcome of terrestrial biological evolution we observe may represent some degree of ″frozen

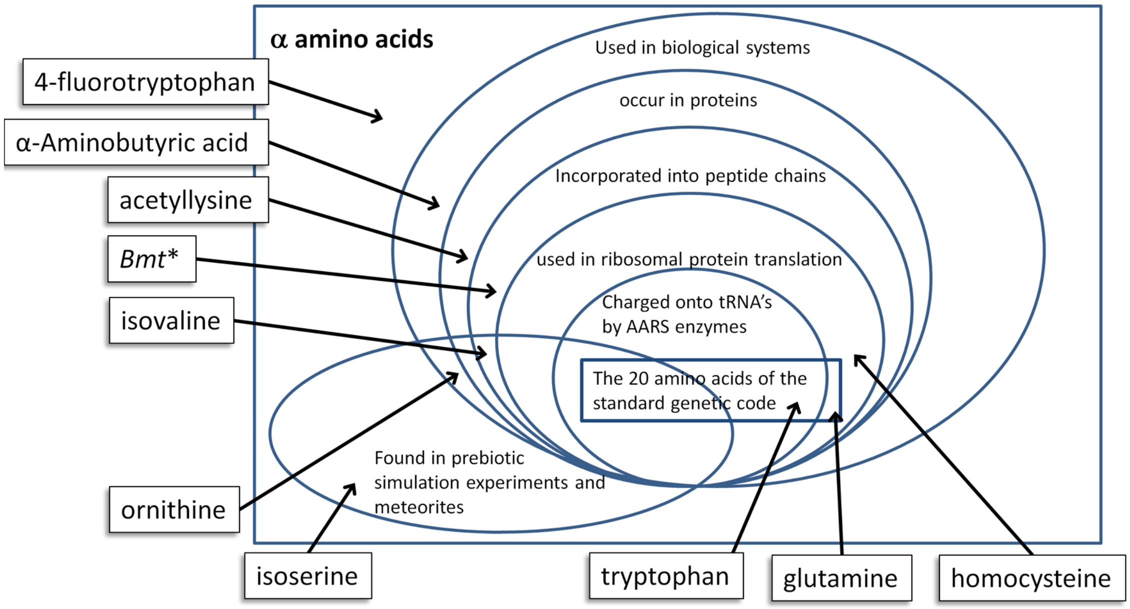


Figure 2. Complex reality of “biological” amino acids visualized as a Venn diagram. The twenty members of the standard amino acid alphabet represent only a small subset of those used in biological systems. Some examples are described in ref 23 (*Bmt is the common abbreviation of 2(S)-amino-3(R)-hydroxy-4(R)-methyl-6(E)-octenoic acid, a product of bacterial nonribosomal peptide synthesis).

accident", as has been advanced for other aspects of the genetic code[28] (see ref 29 for further discussion). Previous attempts to investigate unusual properties of the coded amino acids relative to those of 8 813 isomers derived form the Beilstein database (a set which included an unspecified number of non-α-amino acids)[30] or the set of α-amino acids detected within the Murchison meteorite and a handful of biosynthetic alternatives[31] are based on such small samples that their results should be treated with caution.

A second, overlapping motivation for exploring the α-amino acid universe comes from synthetic biology. Researchers have engineered more than 70 α-amino acids into expanded genetic codes.[32] Given the potential for user-defined coded α-amino acid sets, it is timely to start developing theory about the features of an α-amino acid alphabet that could produce fundamentally new protein structures and functions. An exploration of possible α-amino acid structures forms an important foundation for such work.

A third motivation for considering supersets of α-amino acid structures is more pragmatic. As new instruments and analytical techniques deliver an ever-clearer picture of the true amino acid diversity in extraterrestrial materials, laboratory syntheses, and metabolism, it becomes increasingly useful to generate a map of possible amino acid structures as a foundation for further study. The exact mass of a molecule can be calculated from its structure, and structure-based prediction of chromatographic retention times and mass spectral fragmentation patterns is becoming more robust, all of which will be useful for identifying novel species using hyphenated analytical techniques such as GC-EI/MS.[33,34]

Before these questions can be addressed experimentally, a set of possible amino acid structures must first be generated, the central challenge being the definition of the set(s) of relevance. The combinatorial possibilities inherent to organic structure generation precludes straightforward enumeration of every structure implied by any general definition. With the number of possible relatively low molecular weight isomers numbering conservatively in the millions to trillions, computational methods are likely to provide the only immediate approach. We present here a first exploration of the set of α-amino acid chemical structures of direct relevance to astrobiology and synthetic biology.

■ METHODS

Our aim was to explore the universe of chemical structures implied by the description "α-amino acid", with an emphasis on amino acids of similar size and composition of those of the standard genetic code. The main tool for this study was structure generation software. A structure generator is a computer program that uses a molecular formula as minimal input, from which a library of all possible structural formulas corresponding to this formula is generated as output. Because an α-amino acid includes a carboxyl group and an amino group connected by a substituted methylene linkage (the "α-carbon") (Figure 1), and the substitutions present on the α-carbon can be of almost any size and composition, "α-amino acid" comprises an indefinitely large number of molecules. In other words, the concept of "α-amino acid" is too vague to function as input for any structure generation software. Our challenge was therefore to further restrict the definition of "α-amino acid" so as to provide suitable input for the software to generate a library of the structures relevant to the goals stated in the introduction.

Principles to navigate this challenge of restricting "α-amino acid" come from the long history of using of computer programs to construct all constitutional isomers having the same molecular formula, which is rooted in the earliest development of chemoinformatics[35,36] and research into life's emergence from the abiotic cosmos.[37,38] The representation of chemical compounds as graphs was a foundational step in the application of graph theory,[39] and the efficient, exhaustive generation of these chemical structures has proven a significant challenge ever since.[36,40,41] Although acyclic structures are relatively easy to solve, cyclic structures have proven far more challenging. Lederberg and colleagues solved the problem for many classes of molecules;[37,42] however, limitations in their approach[36] favored its replacement by a more straightforward technique named *orderly generation*.[43−45] This approach imposes an artificial ordering on the graph representation of chemical structure space, thereby allowing the algorithm to recognize early in its execution the intermediates which would lead to duplicate structures. Recent work to adapt orderly generation to molecular graphs[36] has produced software that allows systematic exploration of the universe of amino acid structures for the first time. Structural constraints can be added to the input in order to increase the specificity of results. For example, structural constraints can comprise lists of prescribed or necessary substructures (so-called "good-lists") and forbidden or implausible substructures (so-called "bad-lists".)

We thus used orderly generators to scrutinize the molecular formulas and structural properties of the coded amino acids, deriving good-lists and bad-lists from considerations of coherent structural motifs apparent within this set, as well as motifs which would likely be chemically unstable or implausible based on what is known of organic reactivity in water (see Supporting Information S2). In this way, we defined structure generation tasks with extensive constraints that result in what we believe to be a significant coverage of the structural universe for α-amino acids of potential relevance to biology.

Structure Generation. We used two structure generators to build α-amino acid libraries: MOLGEN 3.5 and MOLGEN 5.

The first, MOLGEN 3.5, is a technically mature software product that underwent its main development phase 15 years ago. MOLGEN 3.5 is endowed with a graphical user interface which supports substructure editing and structural constraint specification. Structural constraints such as ring size and forbidden and prescribed substructures can be defined. However, this software does not allow batch processing of different structure generation tasks, processing just one molecular formula per structure generation task. Full mathematical and algorithmic details of this software are described in ref 46, and a reference guide is offered in ref 47.

The second structure generator used was MOLGEN 5,[48] which was developed more recently (2005−2009). In contrast to its ancestors, MOLGEN 5 is equipped with a command line interface. MOLGEN 5 structure generation tasks are defined similarly, but not equivalently, to those determined by MOLGEN 3.5. For example, within MOLGEN 5 a ring is simply a closed path in the molecular graph, whereas in MOLGEN 3.5 a ring must fulfill the additional condition that no atom is immediately connected to more than two others in the path.

The major benefit of MOLGEN 5 to this effort was the software's ability to process so-called *fuzzy formulas*. Instead of prescribing exact occurrence numbers for the atoms of each chemical element, broader numerical intervals are allowed. For instance, a fuzzy formula that includes all molecular formulas of the 20 coded amino acids would be $C_{2-11}H_{5-14}N_{1-4}O_{2-4}S_{0-1}$. Indeed, this is the *unique* fuzzy formula that describes the least set of molecular formulas which includes all of the coded amino acids.

Another advantage of MOLGEN 5 is its efficient filter for aromatic duplicates, which is embedded in the structure generation process. In MOLGEN 3.5 filtering aromatic duplicates was only available as a postprocessing step, which meant that all generated structures had to be present in the computer's random access memory (RAM) prior to scrutinization, and the range of possible tasks was therefore limited by available memory resources.

Partial Order on Molecular Formulas. Subset relations were important for the design of our libraries; therefore, the relationships between molecular formulas warrant some discussion. Mathematically, a molecular formula f can be considered as a mapping from a set of chemical elements E onto the set of natural numbers. This mapping relates each chemical element X with its multiplicity $f(X)$. For the domain of the coded α-amino acids only five chemical elements have to be considered: $E = \{C, H, N, O, S\}$. As a simple example, $C_2H_5NO_2$ (the molecular formula of glycine) is represented by the mapping f with $f(C) = 2, f(H) = 5, f(N) = 1, f(O) = 2$, and $f(S) = 0$.

Using this representation, we can define a partial order with respect to molecular formulas. We say f_1 is a subformula of f_2, if for all elements $X \in E$ the inequality $f_1(X) \leq f_2(X)$ holds. We use the notation $f_1 \leq f_2$, for instance $C_2H_5NO_2 \leq C_3H_7NO_3$ (the molecular formula for alanine). But note that this does not define a total order on the set of molecular formulas. A total order would require any two formulas f_1, f_2 to be in relation, i.e. $f_1 \leq f_2$ or $f_2 \leq f_1$. As an example, for $C_3H_7NO_3$ and $C_3H_7NO_2S$ (the molecular formula for cysteine), neither of the two inclusions holds.

Regarding fuzzy formulas, this order can be used to describe the set of molecular formulas defined by a fuzzy formula. For instance the fuzzy formula $C_{2-11}H_{5-14}N_{1-4}O_{2-4}S_{0-1}$ includes all molecular formulas f that fulfill the inclusions $C_2H_5NO_2 \leq f \leq C_{11}H_{14}N_4O_4S$.

Backbone and Side Chain. α-Amino acids are composed of a common structural backbone, which is involved in the formation of a linear polypeptide chain, and a variable side chain, which provides the structural variation responsible for determining protein structure and function (Figure 1). In the case of an α-amino acid, the backbone consists of a carboxyl group and an amino group connected by a substituted methylene linkage. Although this linkage could exist at various oxidation states, we treat it here as a simple tetra-substituted junction. We focused on α-amino acids following previous arguments that α-amino acids are the simplest amino acid type capable of forming a repeating structural motif.[19,49] This structural simplicity corresponds to a lack of stereochemical ambiguity, high abundance as products of abiotic synthesis and efficiency of biosynthesis—all of which suggest relevance to biology and genetic coding.[19] Our methodology could of course equally well be applied to any other definable amino acid motif.

All coded amino acids, except proline, have the generic formula $H_2NCH(R)COOH$, where R represents the side chain (structural type Ia in Figure 1).

Proline, represented by generic structural type IVa, is a secondary amine. In other words, the nitrogen atom present in the unionized form of proline is only bound to a single H atom. For the set of coded amino acids, this is the only variation within the backbone. It is worth noting that α,α-dialkyl amino acids (structural type Ic) are quite common in extraterrestrial materials and the products of laboratory syntheses but unknown in coded proteins, though they do appear in nonribosomally coded peptides (see for example ref 50). With the exception of proline, secondary amino acids are not geneticallly coded, and it seems likely that tertiary or quaternary amines (structural types IVb and IVc, respectively) would be incompatible with polypeptide structure, as they would either result in hydrolytically unstable charged peptide bonds (structural type IVb), or be unable to form peptide bonds (structural type IVc).

The side chain of an amino acid allows almost infinite variation, and this is the crux of our study. Focusing on the side chains, we analyzed the coded α-amino acids according to the partial order introduced above. Figure 3 shows a Hasse diagram[51] of the coded amino acids as a *partially ordered set* (or *poset*) using this subset relation. To be more precise, we used hydrogen-reduced formulas of the side chains, which are printed below each amino acid's abbreviation in Figure 3. For glycine, the side chain consists only of one hydrogen atom, thus we have an empty formula below Gly. Amino acids are arranged from bottom to top to reflect an increase in the number of carbon atoms present.

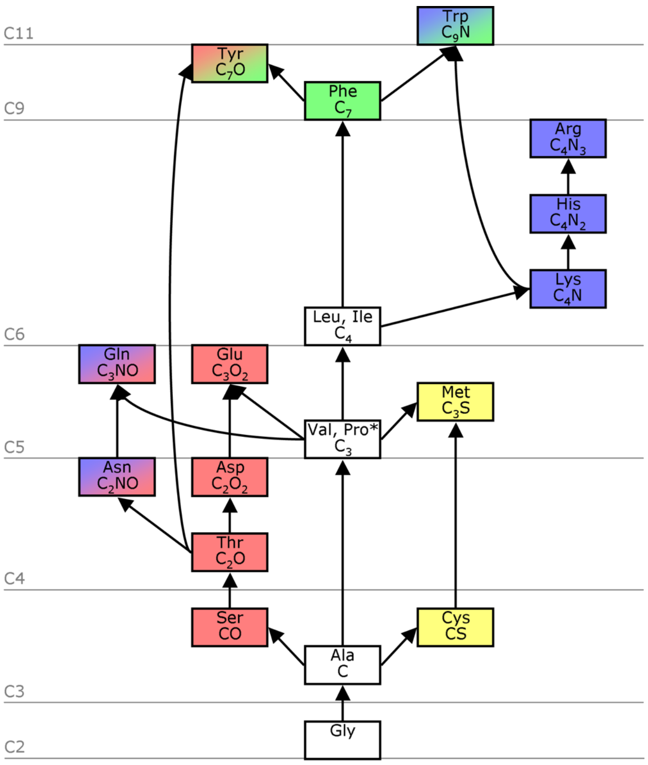


Figure 3. H-reduced formulas of side chains of the 20 genetically encoded α-amino acids represented as posets. Color coding highlights amino acids with heteroatoms and/or a benzene ring. The color codes of these features are red for O, blue for N, yellow for S, and green for a benzene ring.

Classification of the Coded Amino Acids. We classified the 20 genetically encoded amino acids according to six structural properties. Five of these properties are features of the side chain: occurrence of a C, N, O, and S atom, and the presence of a benzene ring. The sixth property refers to whether the backbone is of structural type Ia or IVa (Figure 1). Table 1 shows the considered properties of the coded amino acids.

Table 1. Structural Properties of Side Chains and Backbones of the 20 Coded Amino Acids

	side chain					backbone	
amino acid	C	N	O	S	benzene	Ia	IVa
Gly							
Ala	×					×	
Ser	×		×			×	
Cys	×			×		×	
Thr	×		×			×	
Asp	×		×			×	
Asn	×	×	×			×	
Pro	×						×
Val	×					×	
Met	×		×			×	
Glu	×			×		×	
Gln	×	×	×			×	
Leu	×					×	
Ile	×					×	
Lys	×	×				×	
His	×	×				×	
Arg	×	×				×	
Phe	×				×	×	
Tyr	×		×		×	×	
Trp	×	×			×	×	

On the basis of these properties, we grouped the coded amino acids into 10 classes, with five amino acids each occupying a class of their own:

1. Gly, the only genetically encoded amino acid without any heavy atoms in the side chain;
2. Pro, the only genetically encoded amino acid with a secondary amine in the backbone;
3. Phe, the only genetically encoded amino acid with a benzene ring and no heteroatoms in the side chain;
4. Trp, the only coded amino acid with a benzene ring and N in the side chain;
5. Tyr, the only genetically encoded amino acid with a benzene ring and O in the side chain.

Two classes are composed of two amino acids each:

6. Cys and Met, the only two genetically encoded amino acids with S in the side chain;
7. Asn and Gln, the only two genetically encoded amino acids with N and O in the side chain.

A further class has three members:

8. Lys, His, and Arg, which contain N, but neither O nor a benzene ring in the side chain.

The largest two classes contain four members each:

9. Ala, Val, Ile, Leu, which have no heteroatoms in the side chain, and do not belong to one of the classes listed above;
10. Ser, Thr, Asp, Glu, which contain O, but neither N nor a benzene ring in the side chain.

Figure 4 shows a Venn diagram of this classification.

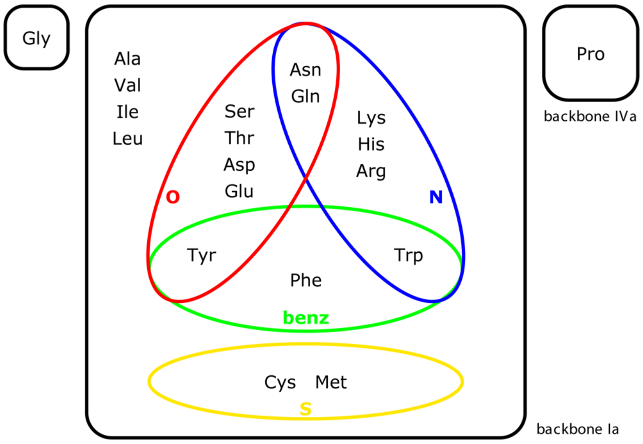


Figure 4. Classification of the 20 genetically encoded amino acids according to the considered properties shown in Table 1

These ten classes were used to define ten structure generation tasks, which were solved by ten computations using MOLGEN 5. While our study is focused on the biologically coded α-amino acids, such a classification can be applied to any given set of input structures and structural properties. For instance, one could add abiotic and biosynthetic amino acids as described in references 25, 27, and 31.

Formulating Generator Input. Using these 10 classes, we then composed fuzzy formulas for each class of side chain as follows:

Heteroatoms. The range should cover exactly the number of heteroatoms observed for the amino acids in each class. For instance the class of Ser, Thr, Asp, and Glu includes O as a side chain heteroatom. There is one O atom in the side chains of Ser and Thr, and two O atoms in the side chains of Asp and Glu. Thus there are 1−2 O atoms for this class.

C atoms. The upper limit is determined by the amino acid with the greatest number of C atoms. The lower limit is chosen to be as small as possible for an amino acid to be a member of the given class. Continuing the example above, Glu's side chain has the maximum number of three C atoms in this class, whereas the lower limit is defined by the fact that a backbone of structural type Ia must be connected to at least one carbon to preclude the propensity for racemization which would accompany the positioning of more electronegative atoms such as N, O or S in this position. Thus, the range for C in this class is 1−3.

H atoms. The maximum is determined by saturated structures, the minimum is chosen as small as possible to generate members of the corresponding class, but not below three (except for Gly). For Ser, Thr, Asp, and Glu, the upper bound of H atoms is seven, calculated via the maximum double bond equivalents of $C_{1-3}O_{1-2}$, and the lower bound is three.

Note that by this procedure the fuzzy formula for each class is uniquely determined. Table 4 shows the fuzzy formulas of the side chains for each of the ten classes. These formulas serve as a major part of the input for the structure generator.

Avoiding Implausible Chemical Structures. Not all molecular structures generated by graph theory methods represent chemically plausible structures.[52] There are some simple graph-theoretical criteria that help to characterize plausible chemical structures, e.g. ref 53. For example, Fink and Reymond[54] neglect structures containing one or more atoms in 3- or 4-membered rings as generally unstable due to ring strain. We accepted and generalized this rule by setting the minimum ring size to 5.

However, such simple criteria are typically insufficient to characterize implausible structures. Structure generators like MOLGEN allow the user to enter lists of forbidden substructures, so-called bad-lists. In the past, several attempts have been made to formulate general bad-lists, for example by Varmuza et al.[55] MOLGEN 5 is shipped with two bad-lists, one with forbidden cyclic and unsaturated substructures and another one containing forbidden bridged aromatic substructures,[48,56] which include structures with disallowed ring strain.

Upon inspecting the remaining generated structures, chemists often recognize further compounds which may be unstable under given chemical conditions. A user-defined bad-list must then be created. This is typically done as depicted in Figure 5: the user iteratively (i) generates structures, (ii) examines the output, (iii) identifies substructures that should be forbidden, (iv) adds these substructures (or any generalization of the problem they represent) to the bad-list, and (v) restarts the generator with the extended bad-list.

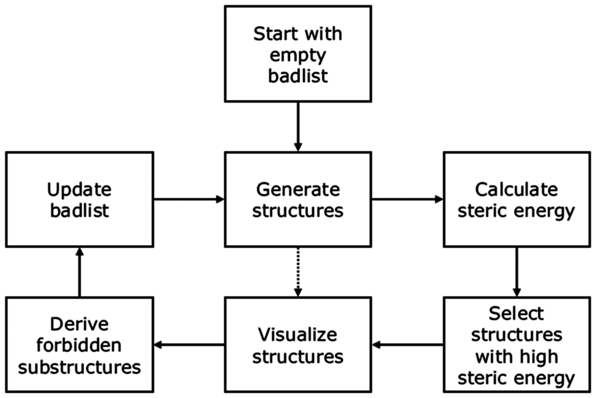


Figure 5. Workflow to obtain a user-defined bad-list.

After several iterations, all unlikely structures can be eliminated in this manner. However, if the output is very large then increasingly rare, implausible structures become increasingly hard to find, by eye and hand, among the staggering number of plausible structures. For the latter situation, we improved the process of building the bad-list as follows: it is known from previous studies that implausible or unstable chemical structures typically have higher steric energy values than plausible or stable ones.[52,57] We therefore used MOLGEN-QSPR[58] to calculate steric energy values based on an MM2 force field method.[59] We then sorted the generated structures according to descending energy values, visually inspected the structures with high energy values and continued the process of extending the bad-list as described above. In this manner a bad-list of 156 forbidden substructures was assembled (see Supporting Information S1). As an unexpected insight, this method made it clear that the atom directly connected to the backbone in structural type Ia (Figure 1) must be C or H. This insight has important ramifications, because larger fixed substructures significantly reduce the combinatorial possibilities for the remainder of the molecule. The effect of systematically removing bad-list entries on a library is shown in Supporting Information S2 and S3. It should be noted that bad lists may at times eliminate structures which might be considered legitimate in some instances, these can thus be made as conservative or liberal as one desires, and some caution must be exercised.

Program Efficiency Considerations. There are two possible ways to introduce backbone type Ia and the neighboring β-C atom as described above. One may either

Table 2. Numbers of Isomers for the 20 Coded α-Amino Acids

		number of isomers		
amino acid	molecular formula	total[a]	no. of 3- and 4-rings	with restricted backbone[b]
Gly	$C_2H_5NO_2$	84	53	1
Ala	$C_3H_7NO_2$	391	244	1
Ser	$C_3H_7NO_3$	1391	857	2
Cys	$C_3H_7NO_2S$	3838	2422	2
Thr	$C_4H_9NO_3$	6836	4242	4
Asp	$C_4H_7NO_4$	65500	25036	14
Asn	$C_4H_8N2O_3$	210267	81702	45
Pro	$C_5H_9NO_2$	22259	8462	3 (6)
Val	$C_5H_{11}NO_2$	6418	3973	2
Met	$C_5H_{11}NO_2S$	86325	54575	10
Glu	$C_5H_9NO_4$	440821	172617	71
Gln	$C_5H_{10}N_2O_3$	1360645	539147	207
Leu, Ile	$C_6H_{13}NO_2$	23946	14866	4
Lys	$C_6H_{14}N_2O_2$	257122	162054	31
His	$C_6H_9N_3O_2$	89502542	13563099	902
Arg	$C_6H_{14}N_4O_2$	88276897	36666235	3563
Phe	$C_9H_{11}NO_2$	277810163	25316848	571 (6)
Tyr	$C_9H_{11}NO_3$	2132674846	209838248	8309 (43)
Trp	$C_{11}H_{12}N_2O_2$	1561538202786	64968283073	559128 (1770)

[a]”Total” isomers denotes all structures which satisfy Lewis electron pairing rules for a given formula, and the following column reports this number of isomers minus those structures which include 3- or 4-membered rings. [b]This is the total number of structures with the given formula which also contain the required backbone motif (see Figure 1). For Pro, the number of isomers shown in parentheses is derived from the assumption that backbone type IVa is used. For the three largest amino acids (Phe, Tyr, and Trp), we show in parentheses the numbers of isomers that have, in addition to backbone Ia, a benzene ring in the side chain.

Table 3. Composition of Unique Libraries of Virtual α-Amino Acids Itemized by the Total Number of C atoms[a]

		number of structures		
number of C atoms	number of formulas	total	plausible	CPU time to compute plausible structures
3	36	5185	5	0.9 s
4	60	202682	88	22.7 s
5	84	4899064	3562	3 m 30 s
6	108	97627979	117389	19 min 7 s
7	132	1776370818	2868117	2 h 19 min 43.6 s
8	156	30987520710	58002850	45 h 27 min
Σ (3−6 C)	288	102734910	121044	23 min 1 s
Σ (3−8 C)	576	32866626438	60992011	48 h 10 min

[a] Processing times refer to MOLGEN 5.01 running under Linux on an Intel Core 2 CPU running at 2.66 GHz.

input the complete substructure as a good-list item, or replace the substructure as *a single* "macro atom". The second method is in some respects more laborious, because the macro atom has to be expanded after structure generation, but the structure generation process in this case turned out to be much faster in terms of computation time when employing this short-cut.

■ RESULTS AND DISCUSSION

We explored three different approaches to defining the universe of potentially codable α-amino acid structures, two of which we used to compute libraries, and only one of which was structurally complete with respect to the coded set. These appraoches were:

1. An exploration of the isomer spaces represented by the formulas of the 20 genetically coded amino acids.
2. A *unique library* (UL) of structures generated by definition of a single fuzzy formula that covers 15 of the 20 coded amino acids, excluding Gly, Pro, and the aromatic α-amino acids Phe, Tyr, and Trp.
3. A *combined library* (CL) that includes all coded amino acids by combining sublibraries obtained from the classification approach described above.

In each case, our emphasis was to characterize the size and broad characteristics of the set of chemical structures: the structures themselves are available as SD files at www.molgen.de/data/AACL.sdf.zip and www.molgen.de/data/AAUL.sdf.zip. Using the free software Open Babel,[60] these files can be converted into various alternative chemical data formats.

Isomer Spaces of Coded Amino Acids. The conceptually simplest and most-easily defined set of chemical structures surrounding the 20 genetically encoded α-amino acids is their isomer space: the exhaustive list of all connectivities of the atoms within each of the 20 coded α-amino acids' side-chains into new chemical structures that are consistent with the fundamental rules of covalent chemical bonding. However, even this concept allows various interpretations, depending on whether or not invariant subunits of structural organization are recognized between the level of individual atoms and whole molecules. For example, the 6-membered aromatic ring in Phe, Tyr, and Trp may be treated as a fixed substructure or as one of many possible cyclic arrangements of six additional carbon atoms. To illustrate how these different interpretations impact the resulting set of generated amino acid structures, we generated isomers according to three different criteria: (i) all isomers without any structural restriction; (ii) isomers without 3- and 4-membered rings; (iii) isomers without 3- and 4-membered rings and with backbone Ia or IVa, respectively.

Table 2 shows the number of chemical structures determined by each of these definitions for each of the coded amino acids. These results were obtained with MOLGEN 3.5 and confirmed by MOLGEN 5.

For Trp, the generation of the 1 561 538 202 786 isomers took 18 days, 17 h, and 47 min (using MOLGEN 3.5, running under Linux on an Intel Core 2 CPU at 2.66 GHz). To our knowledge, this is the largest isomer space ever generated. Computation times for the other cases were shorter by orders of magnitude. Note that all computations were executed on a single core only, i.e. no parallelization was applied. Unless stated otherwise, structures were only generated, but not written to disc. However, writing structures to disc increases computation time by less than 1%.

Beyond the fact that these numbers exceed anything ever suggested in the scientific literature, the clearest trend is a somewhat predictable: generally exponential growth in the number of isomers with increasing number of atoms. However, the number of atoms is not the only factor at work: connectivity of atoms is also important. For instance, there are more isomers for $C_6H_9N_3O_2$ (20 atoms, 11 non-hydrogen atoms, 89 502 542 isomers) than for $C_6H_{14}N_4O_2$ (26 atoms, 12 non-hydrogen atoms, 88 276 897 isomers), because the former has a higher number of double bond equivalents. It is interesting to note that plausible α-amino acids represent ~1 × 10^{-6} to 1.2% of formula structure space given the criteria explored here.

Unique Library (UL) Approach. The isomer spaces of the 20 genetically encoded amino acids are relatively well-defined, but this represents a lower boundary for the size of the universe of amino acid structures we consider useful for the three goals outlined in the introduction, as no known principle dictates that biological or chemical evolution could produce only the chemical structures corresponding to the 20 chemical formulas shown in Table 2. A somewhat more encompassing view comes from consideration of chemical structures that include heteroatoms found within the 20 coded amino acids up to a maximum frequency at which they occur there (i.e., up to 3 nitrogen atoms, 2 oxygen atoms, and 1 sulfur atom in addition to the hydrocarbon framework).

To provide this richer view of the universe of α-amino acid structures, we used MOLGEN 5′s ability to process fuzzy molecular formulas. Specifically, we generated all amino acids containing up to eight C atoms, defined by the fuzzy formula $C_{0-5}H_{3-16}N_{0-3}O_{0-2}S_{0-1}R$ (where R is a trivalent macro atom, representing *backbone type I* plus the β-C atom, for a total of 8 C atoms). We chose a minimum of three H atoms because all coded amino acids with a β-C (i.e., all except except Gly) have at least three H atoms in the side chain.

Table 3 shows the numbers of formulas corresponding to C3−C8 and, in the column labeled *total*, the number of structures generated this way, itemized by the number of carbon atoms. The column labeled *plausible* records the number of structures that remain when we apply stricter restrictions, specifically:

- The two fundamental bad-lists shipped with MOLGEN 5 (cyclic and unsaturated substructures, bridged aromatic substructures, see Methods),
- Our own customized bad-list of 156 substructures deemed likely to be unstable (which we designed to be as inclusive as possible, see Supporting Information S1 for explicit descriptions and rationale),
- Allowed ring sizes of 5−10 (which excludes highly strained 3-and 4-membered rings, but also allows the software to recognize tryptophan as containing a 9-membered ring, counting along the outside of the indole moiety),
- No triple bonds (which do in fact occur in some secondary metabolite amino acids,[22] but which we deemed as too reactive for use in coded proteins).

The subset of 121 044 chemical structures that include up to six carbon atoms encompasses 15 of the coded amino acids (excluding Gly, Pro, Trp, Tyr, and Phe), resulting in an MDL SD file[61] of 149 MB. This library is well suited for further analyses that require a dense coverage of the space of amino acids with backbone Ia, and for which the absence of Phe, Tyr, Trp, and Pro is acceptable. Gly can easily be added back in as it is the only member of its class, but although it is technically possible to extend these calculations to cover the missing portion of the amino acid universe, to do so is likely to be computationally challenging for the foreseeable future. For example, we estimate that calculations for eleven carbon atoms that encompass Trp isomers would produce a library of more than 10^{12} structures (see Figure 6), which exceeds the largest virtual compound library currently available for any type of molecule.[62] This calculation would require massive use of parallel computing. Meanwhile, it is interesting to note that for such simple space as we have charted here, the ratio between total and plausible structures is roughly constant for any given carbon number.

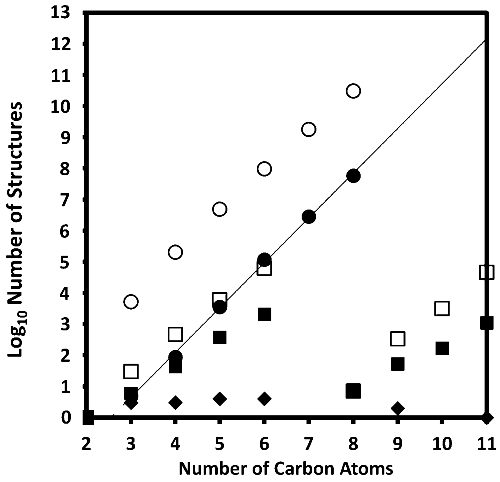


Figure 6. Plot of the number of chemical structures generated as a function of the number of C atoms for the various approaches described. ○ Total structures, UL. ● Total plausible structures, UL. □ Total structures, CL.■ Total plausible structures, CL. ♦ Biologically coded amino acids. The diagonal line is a simple exponential fit to the total UL plausible structure data set, which has a slope of 2 × $10^{-4}e^{3.3144x}$ and an R^2 value of 0.9988. According to this fit we estimate that the UL approach would generate more than 10^{12} structures with eleven C atoms.

It is sobering to note that even were the UL generated exhaustively it would still fail to encompass Pro or any other structures with a backbone of type IVa (Figure 1). These limitations on the results achievable with a single fuzzy formula suggested the use of a third, alternative approach based on generating and combining multiple, nonoverlapping libraries.

Combined Library (CL) Approach. A single "fuzzy formula" cannot be extended to cover all 20 coded α-amino acids without stretching computational power beyond reason for the foreseeable future. A third approach to exploring the universe of amino acid structures avoids these limitations by defining, generating and then combining several nonoverlapping libraries that combine the strongest elements of each previous approach.

Table 4 shows the numbers of structures generated for the ten sublibraries of our classification, each extended into its own fuzzy formula. As in Table 2, we include two numbers for each row, the total number and the number of plausible structures.

Table 4. Definition and Sizes of the Sublibraries of the Combined Library Approach

			no. structures	
amino acid class	sidechain	backbone	total	plausible
Gly	H	Ia	1	1
Pro	$C_{1-3}H_{4-8}$	IVa	38	11
Phe	$C_{6-7}H_{5-7}$	Ia	28	5
Trp	$C_{6-9}H_{6-12}N$	Ia	49296	1307
Tyr	$C_{6-7}H_{5-7}O$	Ia	150	28
Cys, Met	$C_{1-3}H_{3-7}S$	Ia	65	28
Asn, Gln	$C_{1-3}H_{3-7}NO$	Ia	665	97
Lys, His, Arg	$C_{1-4}H_{3-12}N_{1-3}$	Ia	67597	2263
Ala, Val, Ile, Leu	$C_{1-4}H_{3-9}$	Ia	70	22
Ser, Thr, Asp, Glu	$C_{1-3}H_{3-7}O_{1-2}$	Ia	301	84
Σ			118211	3846

Specific treatments were required for the structural classes that encompass the aromatic amino acids, Phe, Trp, and Tyr. Here the presence of a six-membered aromatic ring was enforced using MOLGEN 5′s option to prescribe the number of aromatic bonds, which was set to six (for both runs, generating the total and the plausible structures) and heteroatom aromatics were excluded to reflect the fact that a benzene ring was considered part of the definition of this class.

Another case not covered by the *UL approach* was the class of structures related to Pro. Instead of replacing the backbone (plus the α-carbon atom) by a trivalent macro atom, we simply prescribed backbone IVa as a substructure. Since the formulas are rather simple for this class, this approach did not result in significant problems in efficiency. However, some amino acids known to be abiotic products (such as pipecolic acid) and compatible with protein function[63] were lost. This emphasizes the conservative nature of our combined library approach. Such structures could easily be recovered by extending the class definition, though this would require consideration of amino acids which are not genetically encoded as part of the input definition.

The numbers of structures and the processing times are greatly reduced relative to the UL approach. For the largest sublibrary of 2263 structures the CPU time was 10 s, followed by 8 s for the 1307 compound sublibrary including Trp. All other structure generation processes needed less than a second, thus the construction of the ten libraries was completed in less than half a minute. The combined library of 3846 structures has a size of 4.45 MB as an SDfile. Thus, purely in terms of efficiency, this third approach seems most promising for future investigations.

Figure 6 shows that the combined library contains no structures with seven C atoms. This is because there are no coded amino acids with seven C atoms. One possible way to close this gap without arbitrary changes to our procedure would be to include further amino acids of abiotic or biochemical context into the classification (as mentioned in the case of pipecolic acid, above), for example α-amino-*n*-butyric acid ($C_4H_9NO_2$) or 2-aminopimelic acid ($C_7H_{13}N_1O_4$). The drop-off in the number of structures from six to eight C atoms is caused by the fact that all structures with eight or more C atoms belong to classes where an aromatic ring was prescribed. This constraint narrows the number of combinatorial possibilities by a factor of $\sim 10^5$.

Comparison of the Different Approaches. The plausible structures of the unique library up to six C atoms belong to 214 different molecular formulas, for the plausible structures of the combined library we have only 80 different molecular formulas. For an explicit listing of the formulas, see Supporting Information Table S4. The formulas of the UL cover a mass range from 89.04768 Da ($C_3H_7NO_2$) to 238.07358 Da ($C_6H_{14}N_4O_4S$), those of the CL range from 75.03203 Da ($C_2H_5NO_2$) to 208.12118 Da ($C_{11}H_{16}N_2O_2$). The accurate masses of these molecular formulas could serve as first step toward compound identification in natural samples.

Figure 7 offers a highly abstract look at the parts of the chemical space covered by the three different approaches that we present. Within the figure, black dots represent coded amino acids and shaded areas denote the parts of the amino acid structure universe covered by virtual libraries. The *isomer space* approach (left) is limited to 19 single molecular formulas. Their union includes all structures of the amino acid alphabet and many plausible alternatives, but results in disjointed coverage of the amino acid structure universe. For example, one cannot move from one point to another by a sequence of elementary steps that change the molecular formula by only one or two atoms. The unique library (UL) approach (middle) presents smoother coverage of the possible space, but it cannot cover all given structures with computing resources available presently or in the foreseeable future (this is illustrated by the black dot outside the shaded area which could represent, for example, Pro or Trp). The combined library (CL) approach (right) includes all structures of the genetically encoded amino acids, and much of the coverage of the UL approach. However, it gains considerable speed and efficiency at the price of exhaustive coverage of all of the "interstitial" amino acid structures that lie between the coded amino acids and their close neighbors.

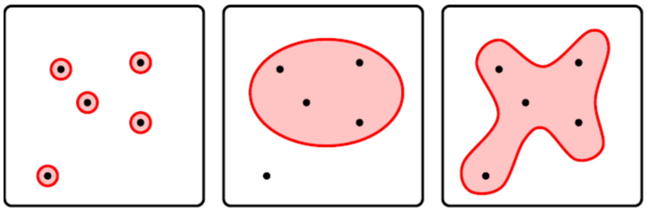


Figure 7. Schematic view of how the different approaches cover the chemical space: isomer space (left), UL (middle), and CL (right); black dots represent coded amino acids.

Even from these simple approaches, interesting and novel insights are clear. For example, we note that 89,051 (>73%) of the structures in the UL contain sulfur. This is partially due to the fact that almost every molecular formula including one sulfur atom has a cognate formula without S, and for formulas with S there are often many more structures than for formulas without S. This is shown in detail in Supporting Information S4, which lists all of the molecular formulas occurring in the libraries of plausible structures, and reports the corresponding number of structures. The overlap of the UL and CL libraries from S4 is 2497 structures.

Comparison with Natural Samples. To check the inclusiveness of our libraries, we compared, using MOLGEN-QSPR, our output CL and UL sets with collated inventories of noncoded biological α-amino acids[22] and α-amino acids found in carbonaceous chondrites.[64] Both natural sets include molecules our libraries *could not* contain, for example because they are α,α-dialkyl-substituted (type Ic, Figure 1), α,*N*-dialkyl-substituted (type IVb, Figure 1), or because they contain elements excluded by our structure generation tasks (see the SI), such as nitrogen or sulfur in excluded oxidation states, phosphorus, 3- or 4-membered rings, or triple bonds.

Twenty-one of the 37 meteoritic α-protonated-α-amino acids reported in carbonaceous chondrites[64] overlap with the CL as do 19 with the UL. However, the 20 coded amino acids, which form the basis of the CL, account for 14 of the 37 described in Burton et al.,[64] as these have long been targets of detection due to their potential importance in the origin of life.[65] The UL does not include glycine, which accounts for the discrepancy between the overlap of the two libraries with the meteoritic set. Neither the CL nor the UL include C7 amino acids, which account for nine of the 37 described in the reference set.[64] The remainder which are absent from both libraries are cyclic or N-alkylated structures.

Forty-seven of the ~700 noncoded biological α-protonated-α-amino acids reported in an extensive review by Hunt[22] were also included in the CL; 87 of these were found in the UL. In contrast with the meteoritic set, the set described by Hunt by definition does not include the coded amino acids.

There are two important implications of these results. First, by extending the library structure generation definitions any number of structures can be generated, underscoring the importance of targeted definition. Second, our libraries produce significant numbers of both naturally occurring biological and abiological α-monoalkyl-substituted-amino acids beyond the coded set, underscoring the utility of this technique in identifying potential natural targets. There may be many as yet undiscovered amino acids in both abiological and biological samples that *are* members of the libraries generated here, which as mentioned above is one of the motivations of this investigation. The extremely restricted and degenerate set of exact masses represented by our libraries' components may offer a significant shortcut to the discovery of novel amino acids in unknown samples.

■ CONCLUSION AND FUTURE DIRECTIONS

We report here three approaches to creating virtual compound libraries that cover the universe of chemical structures surrounding the 20 genetically encoded amino acids. The first of these, simple isomer spaces, are easy to calculate with modern structure generation software, but form a highly conservative lower boundary for any serious investigation of the amino acid structure universe.

The unique library approach covers the chemical space of α-amino acids most densely in terms of molecular formulas, but without vast computing resources it is not possible to generate a library that encompasses all 20 genetically encoded amino acids. The resulting library is estimated to exceed the world's current largest library of chemical structures, a database of drug-like molecules useful for pharmaceutical screening.[62,66] Even this would fail to incorporate well-known amino acids of clear and direct relevance to the fields of astrobiology, synthetic biology, or next-generation instrumentation: it would thus fail to address the research needs outlined in our introduction. Moreoever, incorporation of additional amino acids into the UL approach may lead to a significant increase in the challenge of structure generation and unwieldiness of the results, rendering this approach fundamentally flawed for meaningful exploration of the amino acid structure universe.

Together, these first two approaches underscore our earlier statement that the central challenge for generating a set of possible amino acid structures lies in defining it. Specifically, we advance the view that now and for the foreseeable future no single, meaningful definition of "possible α-amino acids" is worthy of pursuit. Instead, different research emphases must work with different definitions that each explore aspects of a total amino acid structure universe beyond our grasp.

Given this context, a third approach, the combined library, combines the best attributes of each previous definition. It divides the universe of amino acid structures into sublibraries by a mathematically well-defined procedure and thus explores a simple relaxation of preconceptions as to what a prebiotic or genetically encoded amino acid might look like. This approach overcomes the limitations of the isomer space approach by the use of fuzzy formulas to explore beyond the specific details of the 20 genetically encoded amino acids into "nearby" structural territory, and of the unique library approach by sacrificing density of coverage of amino acid chemical space.

This sacrifice is mitigated by the fact that while exhaustive coverage is presently an unachievable goal, posets offer a simple, well-defined and logical framework by which further research can extend our results according to different research needs. In particular, the additive nature of posets removes the need for future investigations to reiterate previously explored localities within the amino acid structure universe. For example, those most interested in prebiotic synthesis and life's origins may simply add to the results presented here those that correspond to additional fuzzy formulas that incorporate other targets of interest.

Viewed in this way, the small size of the combined library (which indicates how conservative it is with respect to applied scientific relevance) can become an asset: this initial compact combined library forms an appropriate platform for future manipulation and analysis, such as the computation of 3-dimensional structures and molecular properties that enable the genetically encoded function of amino acids as building blocks for proteins.

Continuing the theme of interpreting our results as a platform for future research, we reiterate that the present work is based on the constitutional level. This means that tautomeric structures may appear in our libraries. Additional software such as the IUPAC International Chemical Identifier (InChI)[67,68] may be used to identify and remove tautomeric duplicates, reducing the libraries to 89209 (UL) and 3528 (CL) structures (data not shown). However, the application of such a filter illustrates again the importance of definition (and the allied concept of purpose), as in applying such a tautomeric filter, one would need to ensure that tautomers do indeed have equal (or at least very similar) predicted property values, which is not assured in general.[69] This will be the subject of future study.

Our results highlight the deeper point that it is difficult to define the universe of amino acid structures meaningfully without invoking post facto operational definitions. Starting from an observed set of characteristics imposes biases which may or may not be justified or useful from a practical standpoint.

That said, this work is not solely mathematical. Even from the results shown here it is apparent that the set of biologically encoded α-amino acids is extremely small relative to any reasonable definition of plausible structural variation. Undoubtedly, some of the molecules produced by our computations may prove to be problematic in the context of a biological system, whether it be because they are intrinsically unstable under some given set of cytosolic conditions or perhaps incompatible with amino acid charging (for example, in the context of an RNA-driven biochemical system[49]). However, it has been amply documented that additional amino acids can and do enter genetic coding: what we present here is, therefore, a first draft of the chemical space that biochemical systems have available to explore. Why biology, terrestrial or otherwise, "chooses" some isomers and shuns others will be the topic of future explorations. The components of our libraries also present facile targets for in silico evaluation and synthesis for various biochemical applications.[70]

■ ASSOCIATED CONTENT

Ⓢ Supporting Information

S1 forbidden bad-list substructures; S2 effect on the CL output library size of systematically removing each of the 156 bad-list elements; S3 trend in the total number of disallowed structures per number of bad-list entry substructural elements for the CL; S4 numbers of plausible structures per formula for the unique and combined libraries.

This material is available free of charge via the Internet at http://pubs.acs.org.

■ AUTHOR INFORMATION

Corresponding Author

*E-mail: cleaves@ias.edu. Tel.: 858-366-3049.

Notes

The authors declare the following competing financial interest(s): M.M. belongs to the MOLGEN development team which distributes the MOLGEN software for a nominal fee.

■ ACKNOWLEDGMENTS

The authors would like to thank the University of Hawaii's NASA Astrobiology Institute and Institute for Astronomy for generous housing support for M.M. and H.J.C. The authors would also like to thank the NASA Astrobiology Institute's Director's Discretionary Fund for seed funding for this project as part of the NASA Astrobiology Institute under Cooperative Agreement No. NNA09DA77A issued through the Office of Space Science.

■ REFERENCES

(1) Maynard Smith, J.; Szathmary, E. *The Major Transitions in Evolution*; Oxford University Press: New York, 1997.

(2) Des Marais, D. J.; Allamandola, L. J.; Benner, S. A.; Boss, A. P.; Deamer, D.; Falkowski, P. G.; Farmer, J. D.; Hedges, S. B.; Jakosky, B. M.; Knoll, A. H.; Liskowsky, D. R.; Meadows, V. S.; Meyer, M. A.; Pilcher, C. B.; Nealson, K. H.; Spormann, A. M.; Trent, J. D.; Turner, W. W.; Woolf, N. J.; Yorke, H. W. The NASA astrobiology roadmap. *Astrobiology* **2003**, *3* (2), 219−235.

(3) Anfinsen, C. B. Studies on the principles that govern the folding of protein chains. *Science* **1973**, *181*, 223−230.

(4) Canganella, F.; Wiegel, J. Extremophiles: from abyssal to terrestrial ecosystems and possibly beyond. *Naturwissenschaften* **2011**, *98* (4), 253−279.

(5) Cronin, J. R.; Cooper, G. W.; Pizzarello, S. Characteristics and formation of amino acids and hydroxy acids of the Murchison meteorite. *Adv. Space Res.* **1995**, *15* (3), 91−97.

(6) Elsila, J. E.; Glavin, D. P.; Dworkin, J. P. Cometary glycine detected in samples returned by Stardust. *Meteorit. Planet. Sci.* **2009**, *44* (9), 1323−1330.

(7) Miller, S. L. A production of amino acids under possible primitive Earth conditions. *Science* **1953**, *117* (3046), 528−529.

(8) Johnson, A. P.; Cleaves, H. J.; Dworkin, J. P.; Glavin, D. P.; Lazcano, A.; Bada, J. L. The Miller volcanic spark discharge experiment. *Science* **2008**, *322* (5900), 404.

(9) Muñoz-Caro, G. M.; Meierhenrich, U. J.; Schutte, W. A.; Barbier, B.; Arcones Segovia, A.; Rosenbauer, H.; Thiemann, W. H.; Brack, A.; Greenberg, J. M. Amino acids from ultraviolet irradiation of interstellar ice analogues. *Nature* **2002**, *416* (6879), 403−406.

(10) Kuan, Y.-J.; Charnley, S. B.; Huang, H.-C.; Tseng, W.-L.; Kisiel, Z. Interstellar glycine. *Astrophys. J.* **2003**, *593* (2), 848.

(11) Snyder, L. E.; Lovas, F. J.; Hollis, J. M.; Friedel, D. N.; Jewell, P. R.; Remijan, A.; Ilyushin, V. V.; Alekseev, E. A.; Dyubko, S. F. A rigorous attempt to verify interstellar glycine. *Astrophys. J.* **2005**, *619* (2), 914−930.

(12) Belloche, A.; Menten, K. M.; Comito, C.; Mueller, H. S. P.; Schilke, P.; Ott, J.; Thorwirth, S.; Hieret, C. Detection of amino acetonitrile in Sgr B2(N). *Astron. Astrophys.* **2008**, *482*, 179−U137.

(13) Pizzarello, S.; Cooper, G. W.; Flynn, G. J. The nature and distribution of the organic material in carbonaceous chondrites and interplanetary dust particles. In *Meteorites and the Early Solar System II*; Lauretta, D. S.; McSween, H. Y., Eds.; University of Arizona Press in collaboration with Lunar and Planetary Institute: Tucson, 2006; pp 625−651.

(14) Pernot, P.; Carrasco, N.; Thissen, R.; Schmitz-Afonso, I. Tholinomics - Chemical analysis of nitrogen-rich polymers. *Anal. Chem.* **2010**, *82* (4), 1371−1380.

(15) Parker, E. T.; Cleaves, H. J.; Dworkin, J. P.; Glavin, D. P.; Callahan, M.; Aubrey, A.; Lazcano, A.; Bada, J. L. Primordial synthesis of amines and amino acids in a 1958 Miller H_2S-rich spark discharge experiment. *Proc. Natl. Acad. Sci. U.S.A.* **2011**, *108* (14), 5526−5531.

(16) Schmitt-Kopplin, P.; Gabelica, Z.; Gougeon, R. D.; Fekete, A.; Kanawati, B.; Harir, M.; Gebefuegi, I.; Eckel, G.; Hertkorn, N. High molecular diversity of extraterrestrial organic matter in Murchison meteorite revealed 40 years after its fall. *Proc. Natl. Acad. Sci. U.S.A.* **2010**, *107* (7), 2763−2768.

(17) Glavin, D. P.; Callahan, M. P.; Dworkin, J. P.; Elsila, J. E. The effects of parent body processes on amino acids in carbonaceous chondrites. *Meteorit. Planet. Sci.* **2010**, *45* (12), 1948−1972.

(18) Higgs, P. G.; Pudritz, R. E. A thermodynamic basis for prebiotic amino acid synthesis and the nature of the first genetic code. *Astrobiology* **2009**, *9* (5), 483−490.

(19) Cleaves, H. J., 2nd The origin of the biologically coded amino acids. *J. Theor. Biol.* **2010**, *263* (4), 490−498.

(20) Wong, J. T.; Bronskill, P. M. Inadequacy of prebiotic synthesis as origin of proteinous amino acids. *J. Mol. Evol.* **1979**, *13* (2), 115−125.

(21) Yuan, J.; O'Donoghue, P.; Ambrogelly, A.; Gundllapalli, S.; Sherrer, R. L.; Palioura, S.; Simonović, M.; Söll, D. Distinct genetic code expansion strategies for selenocysteine and pyrrolysine are reflected in different aminoacyl-tRNA formation systems. *FEBS Lett.* **2010**, *584* (2), 342−349.

(22) Hunt, S. The Non-Protein Amino Acids. In *The Chemistry and Biochemistry of the Amino Acids*; Barrett, G. C., Ed.; Chapman Hall: London, 1985; pp 55−137.

(23) Freeland, S. "Terrestrial" Amino Acids and their Evolution. In *Amino Acids, Peptides and Proteins in Organic Chemistry*; Wiley-VCH Verlag GmbH & Co. KGaA: Weinheim, Germany, 2009; pp 43−75.

(24) Uy, R.; Wold, F. Posttranslational covalent modification of proteins. *Science* **1977**, *198* (4320), 890−896.

(25) Lu, Y.; Freeland, S. On the evolution of the standard amino-acid alphabet. *Genome Biol.* **2006**, *7* (1), 1−6.

(26) Henze, H. R.; Blair, C. M. The number of structural isomers of the more important types of aliphatic compounds I. *J. Am. Chem. Soc.* **1934**, *56* (1), 157−157.

(27) Lu, Y.; Freeland, S. Testing the potential for computational chemistry to quantify biophysical properties of the non-proteinaceous amino acids. *Astrobiology* **2006**, *6* (4), 606−624.

(28) Crick, F. H. C. The origin of the genetic code. *J. Mol. Biol.* **1968**, *38* (3), 367−379.

(29) Koonin, E. V.; Novozhilov, A. S. Origin and evolution of the genetic code: the universal enigma. *IUBMB Life* **2009**, *61* (2), 99−111.

(30) Zhang, H. Y. Exploring the evolution of standard amino-acid alphabet: when genomics meets thermodynamics. *Biochem. Biophys. Res. Commun.* **2007**, *359* (3), 403−405.

(31) Philip, G. K.; Freeland, S. J. Did evolution select a nonrandom "alphabet" of amino acids? *Astrobiology* **2011**, *11* (3), 235−240.

(32) Liu, C. C.; Schultz, P. G. Adding new chemistries to the genetic code. *Annu. Rev. Biochem.* **2010**, *79*, 413−444.

(33) Schymanski, E. L.; Meringer, M.; Brack, W. Automated strategies to identify compounds on the basis of GC/EI-MS and calculated properties. *Anal. Chem.* **2011**, *83*, 903−912.

(34) Kind, T.; Fiehn, O. Advances in structure elucidation of small molecules using mass spectrometry. *Bioanal. Rev.* **2010**, *2* (1), 23−60.

(35) Gasteiger, J., Engel, T., Eds. *Chemoinformatics*; Wiley-VCH: Weinheim, Germany, 2003.

(36) Meringer, M. Structure enumeration and sampling. In *Handbook of Chemoinformatics Algorithms*; Faulon, J.-L., Bender, A., Eds.; Chapman and Hall: Boca Raton, FL, 2010; pp 233−267.

(37) Lederberg, J. Topological mapping of organic molecules. *Proc. Natl. Acad. Sci. U.S.A.* **1965**, *53* (1), 134−139.

(38) Strick, J. E. Creating a cosmic discipline: The crystallization and consolidation of exobiology, 1957−1973. *J. Hist. Biol.* **2004**, *37*, 131−180.

(39) Trinajstic, N. *Chemical Graph Theory*, 2nd ed.; CRC Press: Boca Raton, FL, 1992.

(40) Faulon, J.-L.; Visco Jr., D.; Roe, D. Enumerating molecules. In *Reviews of Computational Chemistry*; Lipkowitz, K. B., Raima Larter, R., Cundari, T. R., Eds.; Wiley-VCH: New York, 2005; pp 209−286.

(41) Gugisch, R.; Kerber, A.; Laue, R.; Meringer, M.; Rücker, C. History and progress of the generation of structural formulae in chemistry and its applications. *MATCH Commun. Math. Comput. Chem.* **2007**, *58*, 239−280.

(42) Lindsay, R. K.; Buchanan, B. G.; Feigenbaum, E. A.; Lederberg, J. *Applications of Artificial Intelligence for Organic Chemistry: The DENDRAL Project*; McGraw-Hill: New York, 1980.

(43) Faradzhev, I. A. In *Constructive Enumeration of Combinatorial Objects Problèmes Combinatoires et Theorie des Graphes. Colloq. Internat.*; CNRS, University of Orsay, 1976; pp 131−135.

(44) Faradzhev, I. A. Generation of nonisomorphic graphs with a given degree sequence. In *Algorithmic Studies in Combinatorics*; NAUKA: Moscow, USSR, 1978; pp 11−19.

(45) Read, R. C. Everyone a Winner. In *Annual Review Discrete Mathematics*; North-Holland Publishing Company: 1978; Vol. *2*, pp 107−120.

(46) Grund, R. Konstruktion molekularer Graphen mit gegebenen Hybridisierungen und überlappungsfreien Fragmenten. *Bayreuther Math. Schriften* **1995**, *49*, 1−113.

(47) Gugisch, R.; Kerber, A.; Kohnert, A.; Laue, R.; Meringer, M.; Rücker, C.; Wassermann, A. MOLGEN 3.5 Reference Guide. http://molgen.de/documents/molgen35.pdf (accessed August 29th, 2013).

(48) Gugisch, R.; Kerber, A.; Kohnert, A.; Laue, R.; Meringer, M.; Rücker, C.; Wassermann, A. MOLGEN 5.0 Reference Guide. http://molgen.de/documents/manual50.pdf (accessed August 29th, 2013).

(49) Weber, A.; Miller, S. Reasons for the occurrence of the 20 coded protein amino-acids. *J. Mol. Evol.* **1981**, *17* (5), 273−284.

(50) Schwarzer, D.; Finking, R.; Marahiel, M. A. Nonribosomal peptides: From genes to products. *Nat. Prod. Rep.* **2003**, *20* (3), 275−287.

(51) Merrifield, R. E.; Simmons, H. E. *Topological Methods in Chemistry*; Wiley-Interscience: New York, 1989.

(52) Kerber, A.; Laue, R.; Meringer, M.; Rücker, C. Molecules in silico: Potential versus known organic compounds. *MATCH Commun. Math. Comput. Chem.* **2005**, *54*, 301−312.

(53) Rücker, C.; Meringer, M. How many organic compounds are graph-theoretically nonplanar? *MATCH Commun. Math. Comput. Chem.* **2002**, *45*, 153−172.

(54) Fink, T.; Reymond, J.-L. Virtual exploration of the chemical universe up to 11 atoms of C, N, O, F: Assembly of 26.4 million structures (110.9 million stereoisomers) and analysis for new ring systems, stereochemistry, physicochemical properties, compound classes, and drug discovery. *J. Chem. Inf. Model.* **2007**, *47* (2), 342−353.

(55) Varmuza, K.; Jordis, U.; Wolf, G. Database mining for heterocycles: Are structures of small heterocycles generated by a computer program present in databases? http://www.ch.ic.ac.uk/ectoc/echet96/papers/014/ (accessed August 29th, 2013).

(56) Gugisch, R.; Kerber, A.; Kohnert, A.; Laue, R.; Meringer, M.; Rücker, C.; Wassermann, A. MOLGEN 5.0, a molecular structure generator. In *Advances in Mathematical Chemistry*; Basak, S. C., Restrepo, G., Villaveces, J. L., Eds.; Bentham Science Publishers Ltd.: Oak Park, IL, in press.

(57) Schymanski, E. L.; Meringer, M.; Brack, W. Automated strategies to identify compounds on the basis of GC/EI-MS and calculated properties. *Anal. Chem.* **2011**, *83*, 903−912.

(58) Kerber, A.; Laue, R.; Meringer, M.; Rücker, C. MOLGEN−QSPR, a software package for the search of quantitative structure property relationships. *MATCH Commun. Math. Comput. Chem.* **2004**, *51*, 187−204.

(59) Allinger, N. L. Conformational analysis. 130. MM2. A hydrocarbon force field utilizing V1 and V2 torsional terms. *J. Am. Chem. Soc.* **1977**, *99* (25), 8127−8134.

(60) O'Boyle, N. M.; Banck, M.; James, C. A.; Morley, C.; Vandermeersch, T.; Hutchison, G. R. Open Babel: An open chemical toolbox. *J. Cheminf. [Online]* **2011**, 3 (1), No. Article 33, http://www.jcheminf.com/content/3/1/33.

(61) Dalby, A.; Nourse, J. G.; Hounshell, W. D.; Gushurst, A. K. I.; Grier, D. L.; Leland, B. A.; Laufer, J. Description of several chemical structure file formats used by computer programs developed at Molecular Design Limited. *J. Chem. Inf. Comput. Sci.* **1992**, *32* (3), 244−255.

(62) Ruddigkeit, L.; Van Deursen, R.; Blum, L. C.; Reymond, J.-L. Enumeration of 166 billion organic small molecules in the chemical universe database GDB-17. *J. Chem. Inf. Model.* **2012**, *52* (11), 2864−2875.

(63) Anthony-Cahill, S. J.; Magliery, T. J. Expanding the natural repertoire of protein structure and function. *Curr. Pharm. Biotech.* **2002**, 3 (4), 299−315.

(64) Burton, A. S.; Stern, J. C.; Elsila, J. E.; Glavin, D. P.; Dworkin, J. P. Understanding prebiotic chemistry through the analysis of extraterrestrial amino acids and nucleobases in meteorites. *Chem. Soc. Rev.* **2012**, *41* (16), 5459−5472.

(65) Kvenvolden, K.; Lawless, J.; Pering, K.; Peterson, E.; Flores, J.; Ponnamperuma, C.; Kaplan, I.; Moore, C. Evidence for extraterrestrial amino-acids and hydrocarbons in the Murchison meteorite. *Nature* **1970**, *228*, 923−926.

(66) Blum, L. C.; Reymond, J.-L. 970 Million druglike small molecules for virtual screening in the chemical universe database GDB-13. *J. Am. Chem. Soc.* **2009**, *131* (25), 8732−8733.

(67) Heller, S. R.; McNaught, A. D. The IUPAC international chemical identifier (InChI). *Chem. Int.* **2009**, *31* (1), 7−9.

(68) Stein, S. E.; Heller, S. R.; Tchekhovskoi, D. An open standard for chemical structure representation: The IUPAC chemical identifier. *Proceedings of the 2003 International Chemical Information Conference*, Nimes, France, Oct 19−22, 2003; pp 131−143.

(69) Thalheim, T.; Vollmer, A.; Ebert, R.-U.; Kühne, R.; Schüürmann, G. Tautomer identification and tautomer structure generation based on the InChI code. *J. Chem. Inf. Model.* **2010**, *50* (7), 1223−1232.

(70) Duffy, F. J.; Verniere, M.; Devocelle, M.; Bernard, E.; Shields, D. C.; Chubb, A. J. CycloPs: Generating virtual libraries of cyclized and constrained peptides including nonnatural amino acids. *J. Chem. Inf. Model.* **2011**, *51* (4), 829−836.