Journal of Materials Chemistry B

Beyond *DAD*: Proposing a one-letter code for nucleobase-mediated molecular recognition

Journal:	*Journal of Materials Chemistry B*
Manuscript ID	TB-PER-09-2024-001999.R1
Article Type:	Perspective
Date Submitted by the Author:	05-Nov-2024
Complete List of Authors:	Ward, Aiden; University of Rochester, Department of Chemistry Partridge, Benjamin; University of Rochester, Department of Chemistry

Beyond *DAD*: Proposing a one-letter code for nucleobase-mediated molecular recognition

Aiden J. Ward and Benjamin E. Partridge*

Department of Chemistry, University of Rochester, Rochester, NY 14627-0216, United States.

*Email: benjamin.partridge@rochester.edu

Abstract

Nucleobase binding is a fundamental molecular recognition event central to modern biological and bioinspired supramolecular research. Underpinning this recognition is a deceptively simple hydrogen-bonding code, primarily based on the canonical nucleobases in DNA and RNA. Inspired by these biotic structures, chemists and biologists have designed abiotic hydrogen-bonding motifs that can interact with, augment, and reshape native molecular recognition, for applications ranging from genetic code expansion and nucleic acid recognition to supramolecular materials utilizing mono- and bifacial nucleobases. However, as the number of nucleobase-inspired motifs expands, the absence of a standard vocabulary to describe hydrogen bond (HB) patterns has led to a haphazard mixture of shorthand descriptors that are confusing and inconsistent. Alternative notations that specify individual HB sites (such as *DAD* for donor-acceptor-donor) are cumbersome for biological and supramolecular constructs that contain many such patterns. This situation creates a barrier to sharing and interpreting nucleobase-related research across sub-disciplines, hindering collaboration and innovation. In this Perspective, we aim to initiate discourse on this issue by considering what would be needed to formulate a concise one-letter code for the HB patterns associated with synthetic nucleobases. We first summarize some of the issues caused by the current absence of a consistent naming scheme. Subsequently, we discuss some key considerations in designing a coherent naming system. Finally, we leverage chemical rationale and pedagogical mnemonic considerations to propose a succinct and intuitive one-letter code for supramolecular two- and three-HB motifs. We hope that this discussion will spark conversations within our interdisciplinary community, thereby facilitating collaboration and easing communication among researchers engaged in synthetic nucleobase design.

Introduction

The molecular recognition of nucleobases via hydrogen bonds (HBs) is ubiquitous in living systems, enabling the efficient storage, communication, and transfer of genetic information within organisms and between generations. This information is primarily carried in just four patterns of HBs—those of guanine, cytosine, adenine, and thymine/uracil (Figure 1a)—that ensure the fidelity of guanine-cytosine and adenine-thymine/uracil interactions via Watson-Crick-Franklin (WCF) base-pairing. Attracted by the efficiency, specificity, and tunability of nucleobase-mediated molecular recognition, chemists have repurposed DNA for materials design, both as a construction material capable of forming sophisticated architectures via DNA origami (Figure 1b)[1,2] and as a versatile and highly programmable interaction to organize nanoscale matter via colloidal crystal engineering (Figure 1c).[3,4]

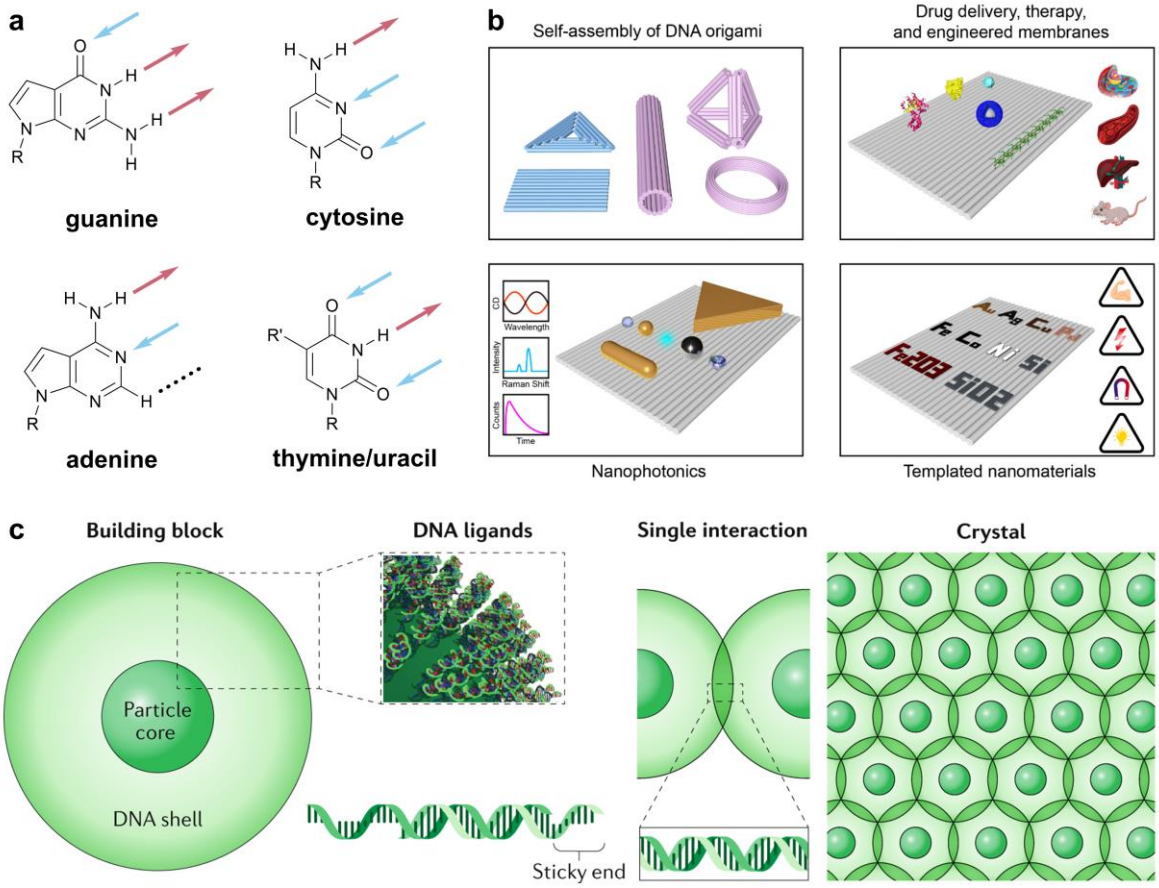


Figure 1. Natural nucleobases in biology and materials. (a) The canonical four nucleobases with hydrogen bonds on the WCF edge denoted by arrows. Color code: blue (HB acceptors) and maroon (HB donors). Black broken line

denotes a null position not involved in WCF base-pairing. (b) Overview of DNA origami-engineered nanomaterials. Adapted from ref. [2]. (c) Overview of colloidal crystal engineering with DNA. Reproduced from ref. [3].

The versatility and potential impact of nucleobase-mediated molecular recognition across chemistry, biology, and materials science have spurred the design of new recognition motifs for different applications, including genetic code expansion, nucleic acid recognition, and supramolecular materials. Genetic code expansion has captivated chemists and biologists alike for decades,[5–10] with recent examples such as hachimoji DNA successfully expanding the genetic alphabet from four to eight nucleobases (hence the name "hachimoji", derived from the Japanese for "eight letters") (Figure 2a).[11] These eight nucleobases rely on six unique HB patterns and the distinct geometries of purine and pyrimidine bases to encode a fully orthogonal four-base pair system that is transcribable *in vitro* using a modified T7 RNA polymerase.[11] In the area of nucleic acid recognition, bifacial nucleobases, known also as Janus nucleobases after the two-faced Roman god of duality,[12] have been designed to present two hydrogen-bonding arrays that each recognize a biotic nucleobase. Seminal work by Lehn and coworkers reported a Janus molecule capable of targeting the cytosine-uracil mismatch.[13] More recently, the groups of Bong[14,15] and Ly[16,17] have utilized the bifacial HB motifs of melamine and Janus nucleobases, respectively, to bind to DNA and RNA double helices and form stable triplexes via double strand invasion (Figure 2b). Triplex formation has also been explored by Rozners and coworkers via the formation of RNA bulges using modified peptide nucleic acids.[18,19] For supramolecular materials, Lehn and coworkers were also instrumental in designing the first synthetic bifacial nucleobase, the G^C motif (Figure 2c),[20] to assemble into hydrogen-bonded macrocycles, the formation of which was confirmed via XRD by Mascal *et al.*[21,22] Subsequent in-depth studies by Fenniri and others explored these rosettes and their nanotubes as fibrous materials and mimics of ion channels.[23–25] The synthesis and assembly of related motifs, including expanded G^C structures,[26–28] A^T Janus nucleobases,[29,30] and a G-C-T triple base-coded nucleobase,[31] have also been reported.

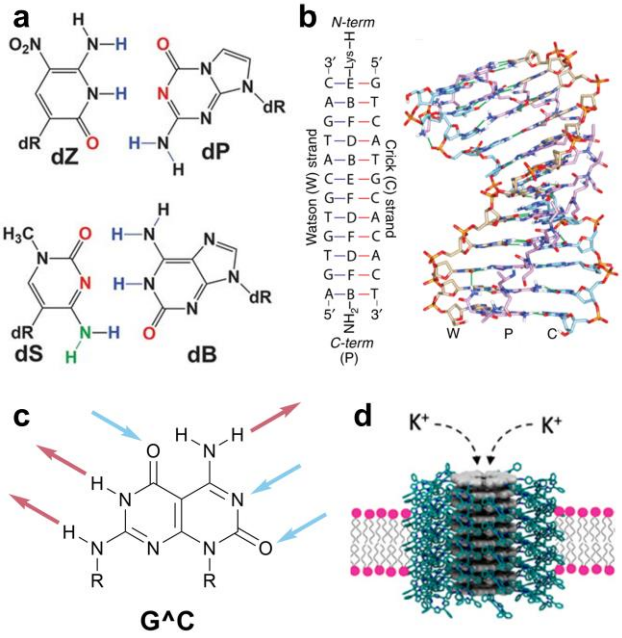
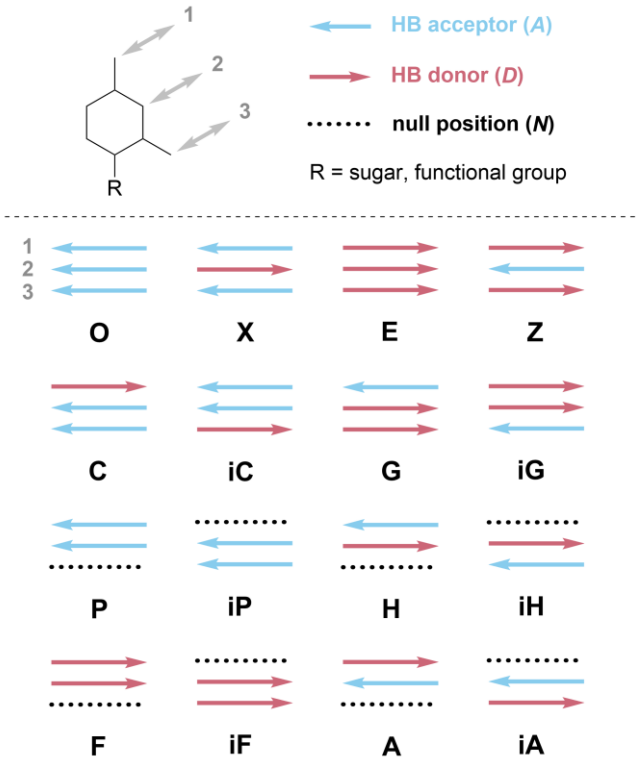


Figure 2. Non-natural nucleobases in biology and materials. (a) Four nucleobases that operate orthogonally to canonical AGCT bases in hachimoji DNA. Adapted from ref. [11]. (b) MD simulation of a DNA-PNA-DNA triplex in which the PNA is modified with Janus nucleobases. Adapted from ref. [17]. (c) The G^C motif.[20] (d) Synthetic rosette nanotube porin mimic assembled from fused G^C bases functionalized with peripheral porphyrin units. Adapted from ref. [25].

Though the development of a rich library of new hydrogen-bonding motifs has been spurred by the transdisciplinary appeal of nucleobase-inspired recognition,[32] new structures have typically been designed and named by separate research communities. These names may have been selected with varying (or absent) rationales to maintain consistency within a specific research group or community, irrespective of usage beyond that community. Such inconsistency stymies collaboration across disciplines. The best alternative at present is to utilize descriptive three-letter codes based on the arrangement of HB donors and acceptors (e.g. *ADD* for guanine, cf. Figure 1a). While this is suitable for single base pairs, these descriptions quickly become cumbersome for multiple, connected HB motifs such as DNA and RNA triplexes or extended supramolecular materials, both enabled by bifacial nucleobases.[17,33] This inefficiency can be analogized to the inconvenience of referring to amino acids by three-letter abbreviations rather than one-letter codes, which becomes infeasible even for small fragment peptides like $A\beta_{16–22}$.[34]

As nucleobase-inspired molecular recognition continues to broaden its relevance in biology and materials science, a succinct, common vocabulary for HB motifs will become ever more necessary to facilitate the community's ability to communicate, compare, and collaborate. Standard vocabularies play a crucial role in chemical and biological research.[35] Three of many examples include the standard one-letter codes for amino acids,[36] the Leontis-Westhof classification of RNA base-pairing pattens,[37] and IUPAC's nucleic acid sequencing notation that not only denotes individual bases (such as G and C) but also all possible combinations thereof (for example, H denotes any base *except* G).[38,39] However, a common shorthand to describe the arrangement of HB sites in nucleobases and related supramolecular building blocks simply does not exist.

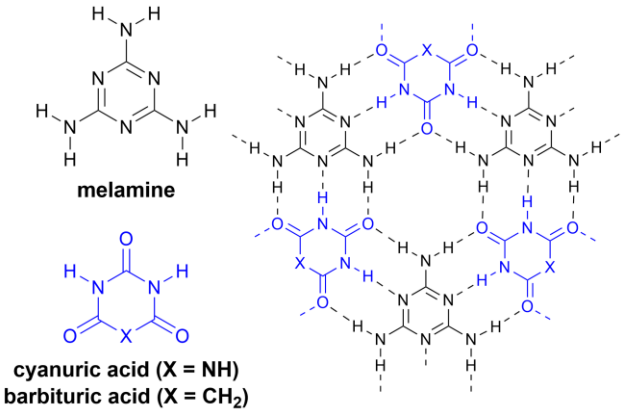


Scheme 1. A possible unified single-letter code for HB patterns. Each of the 16 two- and three-HB motifs can be described as a combination of HB acceptors (*A*, blue arrows), HB donors (*D*, maroon arrows), and non-participating null positions (*N*, black broken lines). HB sites are labelled beginning with the site furthest from any anchoring functional group such as a sugar-phosphate backbone. Rationale for the choice of one-letter codes is discussed later in the main text.

In this Perspective, we seek to initiate discussions to address this need by reflecting on some of the considerations needed to formulate a standard vocabulary for nucleobase-inspired HB motifs. Specifically, we

consider all possible arrangements of three HB positions (acceptors, donors, and null sites) originating from a single molecular edge, comparable to a WCF or Hoogsteen edge; the result is a collection of 16 unique two- and three-HB patterns (Scheme 1). We begin with a brief discussion of the current state of the field, highlighting some of the consequences of uncoordinated naming and missed opportunities that arise therefrom. Subsequently, we outline some considerations for, and inherent contradictions of, designing a cohesive and intuitive naming system. Finally, we conclude by proposing a standard vocabulary and exemplifying its utility. Rather than being a complete solution to this nomenclature challenge, we instead intend that this Perspective act as a starting point to spark discussion around how best to leverage the strengths of the entire community of researchers interested in nucleobase-mediated molecular recognition, for the benefit of all.

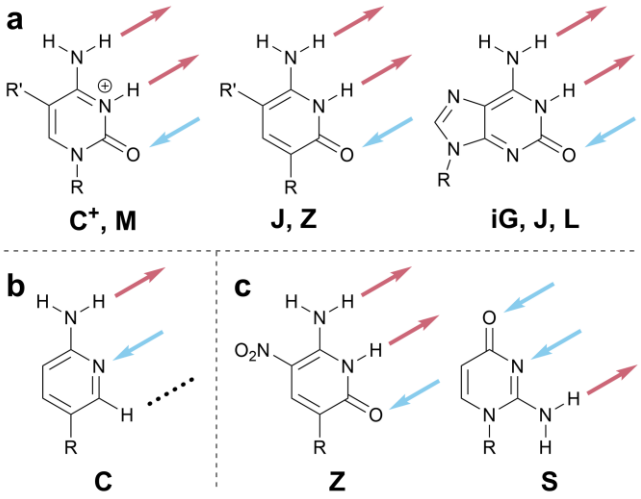
The Consequences of Inconsistent Naming



Scheme 2. Extended two-dimensional HB arrays. Melamine (black) co-assembles with cyanuric acid (blue, X = NH) or barbituric acid (blue, X = CH_2) to form periodic 2D assemblies.

As an example of the current state of nomenclature in this area, we consider the co-assembly of melamine and cyanuric acid (Scheme 2a), extensively studied by Whitesides and coworkers in a series of papers in the early 1990s.[33,40,41] Each edge of melamine features three HB sites, in the order donor-acceptor-donor (abbreviated here in the format *DAD*). Upon mixing with cyanuric acid, which presents the fully complementary acceptor-donor-acceptor (*ADA*) HB pattern, extended two-dimensional (2D) melamine-cyanuric acid (M-CA) structures are formed (Scheme 2, X = NH).[40] However, cyanuric acid is not unique as an *ADA* motif: barbiturates contain the same HB

pattern and also co-assemble with melamine into 2D M-BA structures (Scheme 2, X = CR_2).[42] The *ADA*-capable motif is most commonly found in the nucleobases thymine and uracil, which have three HB sites despite forming only two HBs in their WCF base pairs with adenine (Figure 1a). Indeed, melamine-thymine/uracil co-assembly has been studied in the context of nucleic acid recognition[14,15] and utilized for DNA materials.[43] Describing this recognition using an explicit listing of HB sites is unwieldy: a cyanuric acid-melamine-cyanuric acid triplex would be styled as *ADA^ADA•DAD^DAD•ADA^ADA*, which is sufficiently cumbersome so as to be effectively useless. Moreover, the multiple names used to describe these motifs – M-CA,[40] BA-Mel,[42] M-U,[15] and T-MA-T[43] – obfuscate the HB-mediated molecular recognition event that is common to all of these systems.

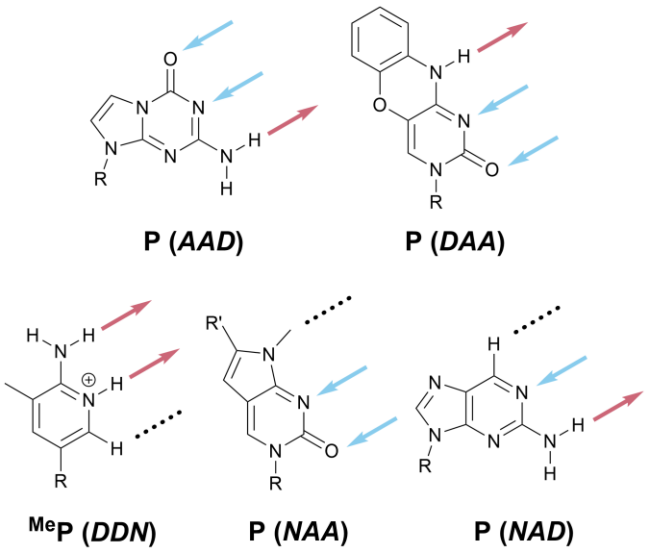


Scheme 3. Uncoordinated naming of nucleobases. (a) The same HB pattern (*DDA*) displayed by three different molecules referred to by six different one-letter codes: C^+,[44–46] M,[45] J,[47,48] Z,[49] iG,[50] and L.[48] (b, c) Hydrogen-bonding patterns named with (b) loose[51] or (c) no[11] rationale.

One pattern, many names. Melamine-cyanuric acid materials exemplify the most common challenge in this space, namely, that no convenient shorthand is associated with a given HB pattern. Consequently, the same HB patterns are referred to using multiple names; this is exacerbated for HB patterns that are not found in WCF base-pairing, such as the *DDA* motif referred to by at least seven different one-letter codes in the literature, including multiple names *for the same molecule* (Scheme 3a). This situation arises because, without a standard rationale by which to select an appropriate shorthand, research groups must select their own one-letter code. Ideally these codes

are chosen to maintain consistency with a group's prior research or to align with the chemical structure of the HB motif (for example, protonated cytosine, C^+, denoting *DDA* in Scheme 3a). However, these shorthands are often based on loose rationale (for example, the cytosine substitute C in Scheme 3b) or are assigned arbitrarily (for example, the hachimoji bases Z and S in Scheme 3c).

One name, many patterns. An additional, arguably more unsatisfactory, outcome of arbitrary naming is that the same one-letter code is applied by different groups to denote different HB patterns. The most common such code is P, which has been employed to denote purines,[50] pyrimidinones,[52–54] pyridines,[52,55] and phenoxazines[56] (Scheme 4). Though such abbreviations are (usually) consistent with a single research group's work, the lack of interoperability between groups and communities hinders collaboration and creates confusion in the literature. In short, the use of one name for many HB patterns, and multiple names for a single HB pattern, are natural consequences of the current naming system for HB motifs *because there is no system*.



Scheme 4. P-bases from the literature. The prevalence of pyridines, purines, and pyrimidines leads to the use of P for multiple different molecules and HB patterns. From left to right: top row, refs [11] and [56]; bottom row, refs [52], [53], and [50].

Considerations for a Cohesive Naming System

The absence of a cohesive and intuitive shorthand reflects the difficulty of assigning nomenclature that is simultaneously satisfactory to multiple research communities, each with different norms, priorities, and

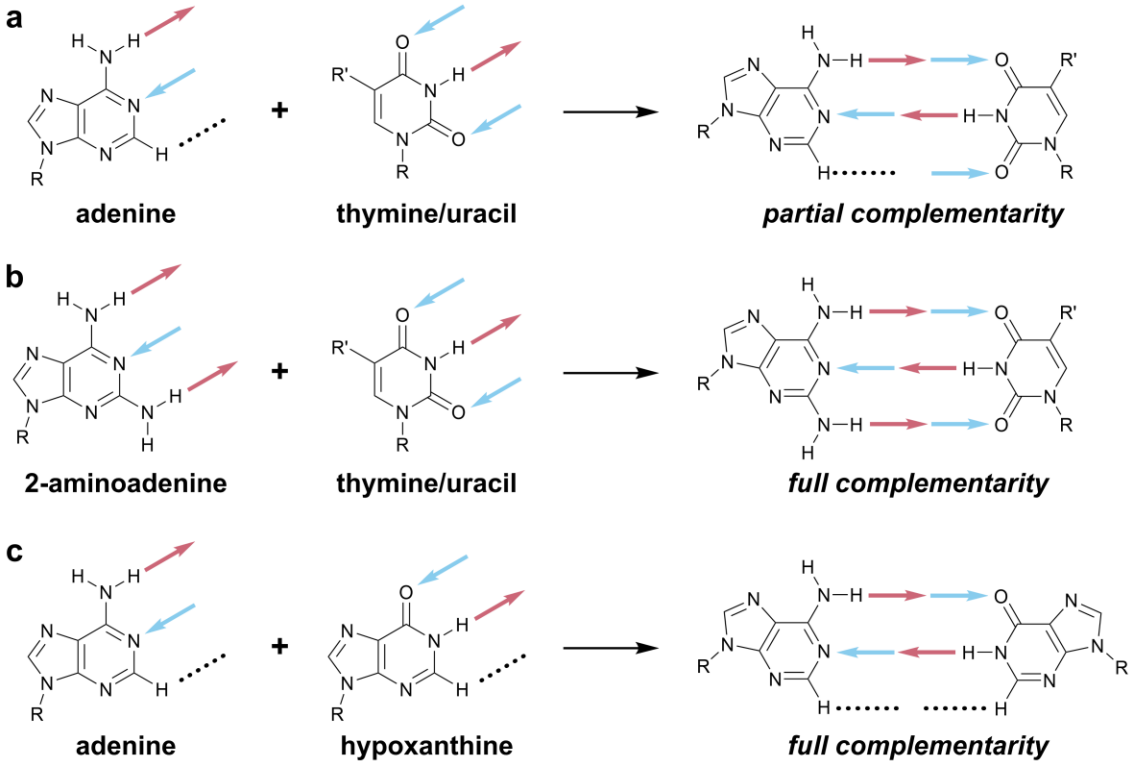
preferences.[57] Scheme 1 presents a potential naming system that describes all possible 2- and 3-HB patterns based on a nucleobase-like scaffold (that is, a six-membered planar heterocycle). In this section, we outline the key factors that we considered in assigning these one-letter codes to specific HB patterns; note that one-letter codes that form part of our proposed vocabulary are styled in bold. We recognize that the list of factors we considered is likely non-exhaustive. Instead, we intend that this scheme serve as a starting point for discussion in the broader research community and iteration towards a universally acceptable solution.

Chemical rationale. As far as possible, the proposed one-letter codes are assigned based on chemical rationale. While this in part reflects the authors' perspective as chemists, we believe that this is a reasonable, evidence-based approach. For example, the most consistently used one-letter codes for HB patterns—G and C for *ADD* and *DAA*, respectively—are those that arise from the nucleosides that utilize these patterns in DNA and RNA, guanosine and cytosine. Assigning a shorthand based on the name of a representative chemical structure that presents that HB pattern is also attractive from a pedagogical perspective,[36,58] to encourage adoption of this vocabulary. Moreover, employing a chemical rationale is consistent with other common naming conventions; for example, the one-letter abbreviations for many of the 20 naturally-occurring amino acids are based on the names of their underlying chemical structures (for example, leucine (L), valine (V), proline (P)).

Consistency with DNA. Perhaps the best recognized associations with specific HB patterns are those associated with the WCF base-pairing of the four nucleobases in DNA and RNA: guanine, cyanine, adenine, and thymine (uracil in RNA), symbolized simply as G, C, A, and T (or U). The ubiquity of these descriptors and their foundation in chemical nomenclature— each is the first letter of the chemical name for the respective nucleobase—favors the assignment of these bases' associated HB patterns with the same single-letter codes. Thus, for example, the HB pattern associated with guanine (*ADD*) is naturally associated with **G**, the HB pattern associated with cyanine (*DAA*) is naturally associated with **C**, and the HB pattern associated with adenine (*DAN*, where N denotes a null site that contains neither a HB donor nor acceptor, Scheme 1) is naturally associated with **A**.

The A–T problem. Inconveniently, thymine frustrates these efforts. Though thymine is complementary to adenine in the context of DNA, the molecular structure of thymine contains a third HB-capable site on the WCF

edge, leading to an *ADA* pattern (Scheme 5a). Though adenine forms a WCF base pair with thymine, the resulting structure utilizes only two of three possible HBs on the WCF edge of the thymine nucleobase (Scheme 5a). Instead, as discussed earlier, a binding partner with a *DAD* pattern is needed to fully complement the HB sites on thymine (Scheme 5b). This contradiction gives rise to what we term the *A–T problem*: though adenine and thymine are complementary in the context of DNA, the perfect complement of thymine in a materials context is not adenine, and vice versa. Therefore, for a standard vocabulary, the one-letter code T would introduce ambiguity as to whether T denotes the complement of A (that is, a 2 hydrogen-bond interaction) or a similar structure to thymine (that is, capable of a three-HB interaction).

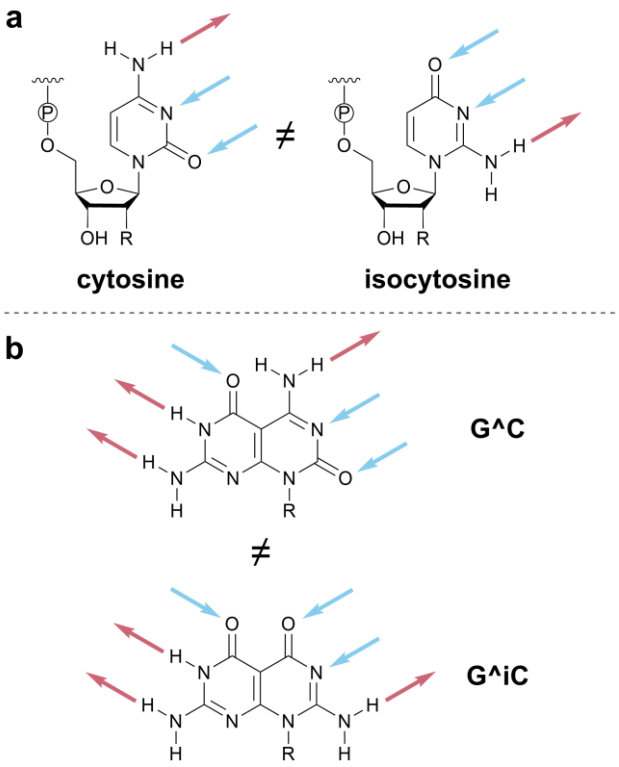


Scheme 5. The A–T problem. (a) The canonical base pair between adenine and thymine/uracil does not fully satisfy all potential HBs on the WCF edge of thymine/uracil. (b) 2-Aminoadenine, known also as 2,6-diaminopurine, is fully complementary to the three HB sites on thymine/uracil. (c) Adenine fully complements the two HB sites on the WCF edge of hypoxanthine.

There are several ways to resolve the A–T problem. One option would be to assign the HB patterns involved in the biological A–T base pair (that is, *DAN* and *ADN*, respectively) with the one-letter codes A and T. In this case, A and T remain complementary but a thymine nucleobase would no longer properly be described as having a T HB

pattern; this is clearly unsatisfactory. An alternative would be to associate A and T with the HB patterns presented by adenine (*DAN*) and thymine (*ADA*), which would maintain the link to these specific molecules but would result in A and T no longer being complementary; this feels counterintuitive and would introduce a clear contradiction with those working with DNA and RNA. Our favored approach avoids the use of T (and, by extension, U) entirely as a descriptor of HB patterns, reserving these letters for the nucleobases themselves, and instead denotes the *ADA* pattern with **X**, derived from the xanthine moiety to which thymine is related. Analogously, the complement of **A**, *ADN*, is denoted **H** for the hypoxanthine motif that exhibits this pattern (Scheme 5c). Though this choice means that the complement of **A** is not T, but **H**, we believe this solution to the A–T problem strikes an appropriate balance between the interests of chemists, biologists, and materials scientists, because it removes a potential source of ambiguity without seeking to redefine a fundamental biological abbreviation (that is, the meaning of the one-letter code, T).

Orientation. Most of the HB patterns in the scope of this Perspective are non-centrosymmetric and therefore cannot be rotated by 180° without changing their presentation of HB donors and acceptors; **C** (*DAA*), for example, is flipped to produce *AAD*, which is no longer complementary to **G** (*ADD*), if the rotation of the **G**-presenting motif is restricted (Scheme 6a). This distinction is salient because most applications of HB-mediated molecular recognition constrain the ability of HB recognition motifs to reorient freely. In DNA, for example, the covalent attachment of the nucleobase to the sugar-phosphate backbone constrains the geometry with which nucleobases can hybridize (Scheme 6a). This effect is valid whenever two or more HB patterns are covalently linked together: consider, for example, the G^C motif (Scheme 6b). Designing an analog in which only one of the HB patterns has been flipped leads to a molecule with molecular recognition properties distinct from those of the parent G^C. Therefore, especially for materials design, distinguishing between two patterns that differ only in their relative orientation is important.



Scheme 6. Importance of HB motif orientation. (a) The *DAA* pattern of cytosine is orientationally related to, but functionally distinct from, the *AAD* pattern of isocytosine. (b) Changing the relative orientation of the two HB patterns in the G^C motif generates a new molecule that will exhibit different supramolecular assemblies.

We invoke the term iso as a directionality descriptor, consistent with literature precedent for the names of isoguanine and isocytosine.[59] Therefore, the pattern for cytosine (*DAA*) is denoted **C** and its rotated analog, *AAD*, is denoted **isoC** or **iC** for short. In assigning the iso designation, the HB pattern should be read from top to bottom when the point of covalent attachment (e.g., to a sugar-phosphate backbone or functional group) is positioned at the bottom of the HB motif (Scheme 6 and top of Scheme 1). Although the use of a modifier strays from a strictly one-letter code, we favored using a modifier to denote the relationship between inverted HB patterns, to minimize the number of unique one-letter codes that would need to be memorized. The modifier 'i' was chosen because alternative modifiers (e.g., asterisk (*), prime (′)) are otherwise defined in nucleic acid notation.

Null in the 2-position. Patterns in which there are only two neighboring HB sites on a single edge (e.g., **A** (*ADN*), where *N* is a null site) can be achieved through molecular design by ensuring that one site cannot participate in a HB, either because that site contains a heteroatom in the wrong geometry to act as a HB acceptor or an X–H

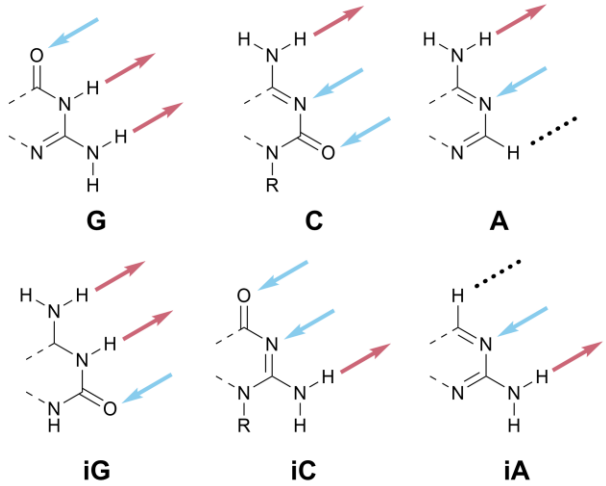
bond (where X is C, N, O) in the wrong geometry to act as a weak HB donor. In contrast, defining a null site in the central, or 2-position, of a 3-HB motif (i.e. *xNx*, where *x* can be any combination of acceptors and donors), is more challenging. The presence of a heteroatom absent an attached hydrogen, possible for the 1- or 3-positions, would instead act as a HB acceptor, and therefore would be described by the appropriate *xAx* pattern. An alternative strategy to define a null 2-position might rely on the presence of a C–H bond in that position. However, abiotic bases that contain a C–H bond in the 2-position, such as difluorobenzene derivatives (Scheme S1a, ESI), are well known to participate in interactions in which the C–H moiety acts as a weak HB donor.[60–62] Furthermore, the presence of a hydrogen atom in this position leads to a steric requirement that the 2-position of the complementary base does not contain a hydrogen atom (Scheme S1b, ESI); accommodating the hydrogen atom of a 'null' C–H bond requires that the complementary base contain a heteroatom, that is, a HB acceptor. It feels natural, therefore, to treat all *xNx* patterns as functionally equivalent to their corresponding *xDx* patterns, even if the strength of the central HB "donor" is negligibly weak, and hence *xNx* patterns are not included in Scheme 1. Although true *xNx* HB motifs are challenging to realize, *xNx* patterns have been assigned one-letter codes for the sake of completeness (Scheme S1c, d, ESI). We note that the extent of donor character, and indeed the electronic properties of nucleobases more broadly, can be assessed through quantum calculations.[63–65] Though these calculations lie beyond the scope of this Perspective, such methods may provide additional guidance for the assignment of a HB pattern to a given structural motif.

A Proposed Naming Scheme for HB Motifs

Scheme 1 at the start of this Perspective outlines our proposed one-letter codes for the 16 unique two- and three-HB patterns of relevance to planar nucleobase-inspired HB motifs. (A complete set of all 27 possible combinations of HB acceptors, HB donors, and null sites is shown in Scheme S2, ESI.) In this section, we introduce each one-letter code and briefly provide specific rationale for its choice.

Canonical bases. The HB patterns associated with the nucleobases guanine, cytosine, and adenine are identified with the same one-letter codes as those bases, that is, **G** denotes *ADD*, **C** denotes *DAA*, and **A** denotes *DAN* (Scheme

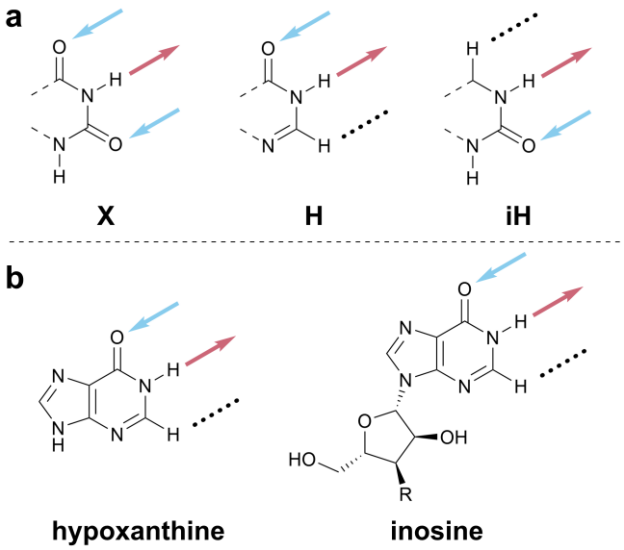
7, top). None of these patterns are centrosymmetric and thus their orientationally-inverted counterparts *DDA*, *AAD*, and *NAD* are denoted **iG**, **iC**, and **iA**, respectively (Scheme 7, bottom).



Scheme 7. HB patterns named according to canonical nucleobases.

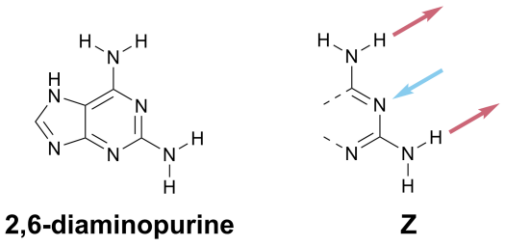
A solution to the A–T problem. To circumvent the A–T problem (*vide supra*), the HB pattern *ADA*, commonly associated with thymine, is denoted **X** for xanthine, and the complement of **A**, *ADN*, is denoted **H** for hypoxanthine (Scheme 8a). **X** is centrosymmetric, and therefore iX is redundant; in contrast, H is non-centrosymmetric and therefore *NDA* is defined as **iH**.

Hypoxanthine is found naturally in tRNA as the nucleobase component of the nucleoside inosine (Scheme 8b); this is a rare example where the names of the nucleobase and nucleoside do not share a common root. We have opted to use hypoxanthine as the source of the name of the *ADN* pattern for two reasons: first, the hypoxanthine motif is more relevant for both biological and materials applications than the full inosine nucleoside; second, using **H** avoids confusion between upper-case I (for inosine) and lower-case i (used here as an abbreviation for *iso*).



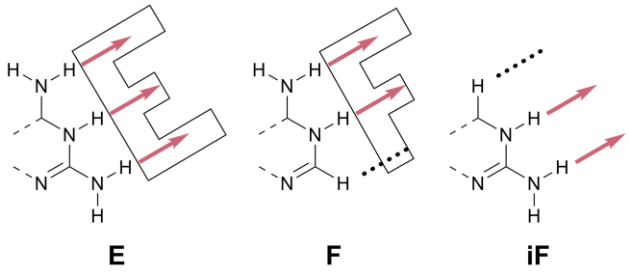
Scheme 8. HB patterns related to a canonical A-T base pair. (a) To avoid the A–T problem, these HB patterns are named according to related chemical structures, xanthine and hypoxanthine (not shown). (b) The *ADN* pattern is denoted as **H** because the nucleobase (hypoxanthine, left) is more structurally relevant for materials than the nucleoside (inosine, right).

Precedence in biology. The perfect complement to **X** is the HB pattern *DAD*, which has been variously described in the literature as A′ (for adenine substitute),[66] K (with no clear rationale),[50,58,67] and D (for 2,6-diaminopurine, a common chemical motif used to obtain the *DAD* motif; Scheme 9).[50,68,69] 2,6-Diaminopurine is also known as 2-aminoadenine, and can be considered as a variant of adenine in which an extra HB donor (the amino group) has been added to achieve full complementarity with all three HB sites of thymine (Scheme 5b). This nucleobase is well established in biological contexts, where it is referred to as the Z-base, and is found in place of adenine in the genomes of some bacteriophages to evade degradation by restriction enzymes.[70–72] To follow this precedent, we propose that **Z** denote the HB pattern *DAD*, such that **X** and **Z** are perfectly complementary.



Scheme 9. The Z HB pattern. (Left) 2,6-Diaminopurine, known also as 2-aminoadenine, is referred to as the Z base in bacterial virus genomes. (Right) The donor-acceptor-donor *DAD* pattern denoted by **Z**.

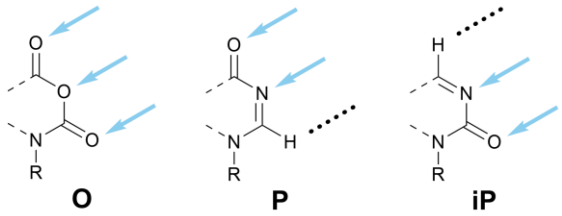
Molecular shape. Other than **G** (*ADD*) and **iG** (*DDA*), two other HB patterns contain two or more adjacent donor sites: *DDD* and *DDN*. To the best of our knowledge, *DDD* has not been explored as a HB motif in nucleobase-like structures, and therefore there is no naming precedent to follow. In contrast, *DDN* has been reported and denoted either as P,[52] which is used to describe many different structures (Scheme 4), or M^+,[73,74] for *am*inopyridine. Given that one consideration for a successful systematic vocabulary is to make it easy to adopt and remember, we propose naming the *DDD* and *DDN* patterns as **E** and **F**, respectively, due to the physical arrangement of HB donors in these motifs (Scheme 10). This rationale mirrors the well-accepted justification for the amino acid tryptophan being denoted W, due to its molecular shape and double-ring structure.[36] **E** is centrosymmetric, but **F** is not; therefore *NDD* is defined as **iF**.



Scheme 10. The E, F, and iF HB patterns. Analogous to the double-ring structure of tryptophan (one-letter code: W), the names **E** and **F** are mnemonic shorthands based on molecular shape.

Other patterns. Three patterns remain unnamed. *AAA*, the complement of **E**, has not been reported, likely due to the challenge of synthesizing a molecule with a stable *AAA*-presenting tautomer. One route to this motif would utilize an oxygen atom as the acceptor in the 2-position (see, for example, the oxazine-dione in Scheme 11). Based on this oxazine core, we denote *AAA* as **O**. Other *AAA*-presenting molecular motifs that utilize halogens as HB acceptors can be envisioned, such as 2,6-difluoropyridine, but such motifs may bind weakly to typical HB donors such as amines due to the relative weakness of organic fluorine as an HB acceptor.[75]

In the previous subsection, we assigned consecutive letters to *DDD* (**E**) and its analog without a HB site in the 3-position, *DDN* (**F**). Analogously, given that *AAA* is denoted **O**, we propose that its analog without a HB site in the 3-position, *AAN*, be denoted **P** (Scheme 11). **P** is non-centrosymmetric and therefore the final unnamed pattern, *NDD*, is denoted **iP**.

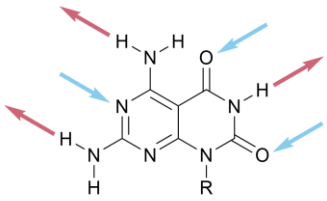


Scheme 11. The O, P, and iP HB patterns. The *AAA* pattern is denoted **O** due to the oxazine-dione motif proposed to define three neighboring HB acceptors. Removing one of these three identical sites generates **P**, named by analogy to the three- and two-donor motifs **E** and **F**.

A Standard Vocabulary in Practice

The one-letter code proposed in Scheme 1 and discussed in the previous section is succinct and unambiguous. The naming of these 16 HB patterns naturally leads to eight base pairs that are fully complementary to each other: **G-C**, **iG-iC**, **A-H**, **iA-iH**, **F-P**, **iF-iP**, **X-Z**, and **O-E**. These base pairs concisely convey information about the nature and number of HBs involved in each base-pairing interaction.

The value of one-letter codes is especially evident when discussing supramolecular systems that invoke more than one HB molecular recognition event. Consider, for example, the cyanuric acid-melamine-cyanuric acid triplex discussed earlier (Scheme 2). Using our proposed vocabulary, the HB pattern of melamine is described as **Z^Z**, where the caret (^) is used in the bifacial nucleobase community to denote the HB patterns on each face of the molecule (see, for example, **G^C** in Figure 2c). Similarly, the HB pattern of cyanuric acid is described as **X^X**. Hence the triplex noted above could be concisely represented as **X^X•Z^Z•X^X**, rather than the *ADA^ADA•DAD^DAD•ADA^ADA* description required by bond-by-bond donor-acceptor notation.



Scheme 12. A Janus-AT (J_{AT}) bifacial nucleobase reported by Asadi *et al.*[29]

It is important to clarify that the proposed naming scheme is not intended to supersede efforts to name HB-capable molecules with trivial or more convenient names. We recognize, for example, that referring to melamine

only as **Z^Z** would be non-sensical, as valuable chemical information about the molecular structure that presents the **Z^Z** HB pattern would be lost. Instead, our vocabulary is meant to augment the current literature and provide a means of communicating HB patterns, rather than molecular structures, clearly and unambiguously. As an example, consider the Janus-AT (J_{AT}) bifacial nucleobase shown in Scheme 12.[29] Under our proposed standard vocabulary, the J_{AT} motif would be denoted by **X^Z**. While this may be less practical for the goals of the original authors' work, including this formalized name makes the connection between the J_{AT} motif and other HB motifs, such as the cyanuric acid (**X^X**) and melamine (**Z^Z**) system, much more readily apparent. Our proposed naming scheme maps readily onto reported bifacial nucleobases (Scheme S3, ESI). Therefore, we encourage researchers to include this formal description of their hydrogen bonding motifs in their future work, to facilitate collaboration and the exchange of ideas within our community.

Summary and Outlook

In summary, we have identified a need for a consistent naming scheme for nucleobase-like HB patterns and proposed a standard vocabulary to address this need. As far as possible, literature precedent, chemical rationale, and consistency with DNA have been considered while designating one-letter codes. Otherwise, mnemonics based on shape or structural relationships have been employed, specifically to assist with memorization and encourage uptake of our proposed alphabet by researchers.

We recognize that a vocabulary is only useful if it is adopted broadly and meets consensus in its intended community of use. Therefore, this Perspective is intended as a starting point, both to solicit feedback and to spur conversations among researchers in our field, to consider and debate how best we can communicate and share our ideas and findings. In this way, we hope that we can contribute to maximizing our collective progress towards our respective goals, whether in the biological, chemical, or materials context, applied or fundamental. With this in mind, we encourage feedback and commentary, to help shape a common vocabulary useful to all.

Author Contributions

A.J.W.: conceptualization, funding acquisition, investigation, visualization, writing – review & editing; B.E.P.: conceptualization, funding acquisition, supervision, visualization, writing – original draft, review & editing.

Data Availability

No primary research results, software or code have been included and no new data were generated or analyzed as part of this Perspective.

Conflicts of Interest

There are no conflicts of interest to declare.

Acknowledgements

This work was supported by the Air Force Office of Scientific Research (AFOSR) through the Young Investigator Program (Award FA9550-24-1-0104). A.J.W. acknowledges the American Chemical Society (ACS) Division of Organic Chemistry for financial support through a Summer Undergraduate Research Fellowship (SURF). The authors thank Parbhat Kumar and Alejandro Lazaro (University of Rochester) for scientific discussions.

References

1) Y. Tian, J. R. Lhermitte, L. Bai, T. Vo, H. L. Xin, H. Li, R. Li, M. Fukuto, K. G. Yager, J. S. Kahn, Y. Xiong, B. Minevich, S. K. Kumar and O. Gang, Ordered three-dimensional nanomaterials using DNA-prescribed and valence-controlled material voxels, *Nat. Mater.*, 2020, **19**, 789–796.

2) P. Zhan, A. Peil, Q. Jiang, D. Wang, S. Mousavi, Q. Xiong, Q. Shen, Y. Shang, B. Ding, C. Lin, Y. Ke and N. Liu, Recent Advances in DNA Origami-Engineered Nanomaterials and Applications, *Chem. Rev.*, 2023, **123**, 3976–4050.

3) C. R. Laramy, M. N. O'Brien and C. A. Mirkin, Crystal engineering with DNA, *Nat. Rev. Mater.*, 2019, **4**, 201–224.

4) D. Samanta, W. Zhou, S. B. Ebrahimi, S. H. Petrosko and C. A. Mirkin, Programmable Matter: The Nanoparticle Atom and DNA Bond, *Adv. Mater.*, 2022, **34**, 2107875.

5) D. A. Malyshev, K. Dhami, T. Lavergne, T. Chen, N. Dai, J. M. Foster, I. R. Corrêa and F. E. Romesberg, A semi-synthetic organism with an expanded genetic alphabet, *Nature*, 2014, **509**, 385–388.

6) J. W. Chin, Expanding and reprogramming the genetic code, *Nature*, 2017, **550**, 53–60.

7) M. A. Shandell, Z. Tan and V. W. Cornish, Genetic Code Expansion: A Brief History and Perspective, *Biochemistry*, 2021, **60**, 3455–3469.

8) M. Kimoto and I. Hirao, Genetic alphabet expansion technology by creating unnatural base pairs, *Chem. Soc. Rev.*, 2020, **49**, 7602–7626.

9) T. Yoshida, K. Morihiro, Y. Naito, A. Mikami, Y. Kasahara, T. Inoue and S. Obika, Identification of nucleobase chemical modifications that reduce the hepatotoxicity of gapmer antisense oligonucleotides, *Nucleic Acids Res.*, 2022, **50**, 7224–7234.

10) A. Berdis, Nucleobase-modified nucleosides and nucleotides: Applications in biochemistry, synthetic biology, and drug discovery, *Front. Chem.*, 2022, **10**, 1051525.

11) S. Hoshika, N. A. Leal, M.-J. Kim, M.-S. Kim, N. B. Karalkar, H.-J. Kim, A. M. Bates, N. E. Watkins, H. A. SantaLucia, A. J. Meyer, S. DasGupta, J. A. Piccirilli, A. D. Ellington, J. SantaLucia, M. M. Georgiadis and S. A. Benner, Hachimoji DNA and RNA: A genetic system with eight building blocks, *Science*, 2019, **363**, 884–887.

12) S. J. Cristol and D. C. Lewis, Bridged Polycyclic Compounds. XLV. Synthesis and Some Properties of 5,5a,6,11,11a,12-Hexahydro-5,12:6,11-di-o-benzenonaphthacene (Janusene), *J. Am. Chem. Soc.*, 1967, **89**, 1476–1483.

13) N. Branda, G. Kurz and J.-M. Lehn, JANUS WEDGES: a new approach towards nucleobase-pair recognition, *Chem. Commun.*, 1996, 2443.

14) Y. Zeng, Y. Pratumyot, X. Piao and D. Bong, Discrete Assembly of Synthetic Peptide–DNA Triplex Structures from Polyvalent Melamine–Thymine Bifacial Recognition, *J. Am. Chem. Soc.*, 2012, **134**, 832–835.

15) S. Rundell, O. Munyaradzi and D. Bong, Enhanced Triplex Hybridization of DNA and RNA via Syndiotactic Side Chain Presentation in Minimal bPNAs, *Biochemistry*, 2022, **61**, 85–91.

16) S. A. Thadke, J. D. R. Perera, V. M. Hridya, K. Bhatt, A. Y. Shaikh, W.-C. Hsieh, M. Chen, C. Gayathri, R. R. Gil, G. S. Rule, A. Mukherjee, C. A. Thornton and D. H. Ly, Design of Bivalent Nucleic Acid Ligands for Recognition of RNA-Repeated Expansion Associated with Huntington's Disease, *Biochemistry*, 2018, **57**, 2094–2108.

17) S. A. Thadke, V. M. Hridya, J. D. R. Perera, R. R. Gil, A. Mukherjee and D. H. Ly, Shape selective bifacial recognition of double helical DNA, *Commun. Chem.*, 2018, 79.

18) T. Endoh, N. Brodyagin, D. Hnedzko, N. Sugimoto and E. Rozners, Triple-Helical Binding of Peptide Nucleic Acid Inhibits Maturation of Endogenous MicroRNA-197, *ACS Chem. Biol.*, 2021, **16**, 1147–1151.

19) C. A. Ryan, M. M. Rahman, V. Kumar and E. Rozners, Triplex-Forming Peptide Nucleic Acid Controls Dynamic Conformations of RNA Bulges, *J. Am. Chem. Soc.*, 2023, **145**, 10497–10504.

20) A. Marsh, M. Silvestri and J.-M. Lehn, Self-complementary hydrogen bonding heterocycles designed for the enforced self-assembly into supramolecular macrocycles, *Chem. Commun.*, 1996, 1527.

21) M. Mascal, N. M. Hext, R. Warmuth, M. H. Moore and J. P. Turkenburg, Programming a Hydrogen-Bonding Code for the Specific Generation of a Supermacrocycle, *Angew. Chem. Int. Ed. Engl.*, 1996, **35**, 2204–2206.

22) M. Mascal, N. M. Hext, R. Warmuth, J. R. Arnall-Culliford, M. H. Moore and J. P. Turkenburg, The G−C DNA Base Hybrid: Synthesis, Self-Organization and Structural Analysis, *J. Org. Chem.*, 1999, **64**, 8479–8484.

23) R. L. Beingessner, Y. Fan and H. Fenniri, Molecular and supramolecular chemistry of rosette nanotubes, *RSC Adv.*, 2016, **6**, 75820–75838.

24) H. Fenniri, G. Tikhomirov, D. H. Brouwer, S. Bouatra, M. El Bakkari, Z. Yan, J.-Y. Cho and T. Yamazaki, High Field Solid-State NMR Spectroscopy Investigation of ^{15}N-Labeled Rosette Nanotubes: Hydrogen Bond Network and Channel-Bound Water., *J. Am. Chem. Soc.*, 2016, **138**, 6115–6118.

25) P. Tripathi, L. Shuai, H. Joshi, H. Yamazaki, W. H. Fowle, A. Aksimentiev, H. Fenniri and M. Wanunu, Rosette Nanotube Porins as Ion Selective Transporters and Single-Molecule Sensors, *J. Am. Chem. Soc.*, 2020, **142**, 1680–1685.

26) A. Asadi, B. O. Patrick and D. M. Perrin, G^C Quartet — A DNA-Inspired Janus-GC Heterocycle: Synthesis, Structural Analysis, and Self-Organization, *J. Am. Chem. Soc.*, 2008, **130**, 12860–12861.

27) G. Borzsonyi, R. S. Johnson, A. J. Myles, J.-Y. Cho, T. Yamazaki, R. L. Beingessner, A. Kovalenko and H. Fenniri, Rosette nanotubes with 1.4 nm inner diameter from a tricyclic variant of the Lehn–Mascal G∧C base, *Chem. Commun.*, 2010, **46**, 6527.

28) G. Borzsonyi, A. Alsbaiee, R. L. Beingessner and H. Fenniri, Synthesis of a Tetracyclic G∧C Scaffold for the Assembly of Rosette Nanotubes with 1.7 nm Inner Diameter, *J. Org. Chem.*, 2010, **75**, 7233–7239.

29) A. Asadi, B. O. Patrick and D. M. Perrin, Janus-AT Bases: Synthesis, Self-Assembly, and Solid State Structures, *J. Org. Chem.*, 2007, **72**, 466–475.

30) D. Shin and Y. Tor, Bifacial Nucleoside as a Surrogate for Both T and A in Duplex DNA, *J. Am. Chem. Soc.*, 2011, **133**, 6926–6929.

31) M. A. Wagh, R. Maity, R. J. Bhosale, D. Semwal, S. Tothadi, R. Vaidhyanathan and G. J. Sanjayan, Three in One: Triple G-C-T Base-Coded Brahma Nucleobase Amino Acid: Synthesis, Peptide Formation, and Structural Features., *J. Org. Chem.*, 2021, **86**, 15689–15694.

32) M. López-Tena, S.-K. Chen and N. Winssinger, Supernatural: Artificial Nucleobases and Backbones to Program Hybridization-Based Assemblies and Circuits, *Bioconjug. Chem.*, 2023, **34**, 111–123.

33) J. A. Zerkowski, C. T. Seto and G. M. Whitesides, Solid-state structures of rosette and crinkled tape motifs derived from the cyanuric acid melamine lattice, *J. Am. Chem. Soc.*, 1992, **114**, 5473–5475.

34) E. R. Piedmont, E. E. Christensen, T. D. Krauss and B. E. Partridge, Amphiphilic dendrons as supramolecular holdase chaperones, *RSC Chem. Biol.*, 2023, **4**, 754–759.

35) N. B. Leontis, R. B. Altman, H. M. Berman, S. E. Brenner, J. W. Brown, D. R. Engelke, S. C. Harvey, S. R. Holbrook, F. Jossinet, S. E. Lewis, F. Major, D. H. Mathews, J. S. Richardson, J. R. Williamson and E. Westhof, The RNA Ontology Consortium: An open invitation to the RNA community, *RNA*, 2006, **12**, 533–541.

36) M. Saffran, Amino acid names and parlor games: from trivial names to a one-letter code, amino acid names have strained students' memories. Is a more rational nomenclature possible?, *Biochem. Educ.*, 1998, **26**, 116–118.

37) N. B. Leontis and E. Westhof, Geometric nomenclature and classification of RNA base pairs, *RNA*, 2001, **7**, 499–512.

38) A. Cornish-Bowden, Nomenclature for incompletely specified bases in nucleic acid sequences: rcommendations 1984, *Nucleic Acids Res.*, 1985, **13**, 3021–3030.

39) Nomenclature committee of the International Union of Biochemistry (NC-IUB). Nomenclature for incompletely specified bases in nucleic acid sequences. Recommendations 1984., *J. Biol. Chem.*, 1986, **261**, 13–17.

40) C. T. Seto and G. M. Whitesides, Self-assembly based on the cyanuric acid-melamine lattice, *J. Am. Chem. Soc.*, 1990, **112**, 6409–6411.

41) J. A. Zerkowski and G. M. Whitesides, Steric Control of secondary, solid-state architecture in 1:1 complexes of melamines and barbiturates that crystallize as crinkled tapes, *J. Am. Chem. Soc.*, 1994, **116**, 4298–4304.

42) T. P. Roche, P. J. Nedumpurath, S. C. Karunakaran, G. B. Schuster and N. V. Hud, One-Pot Formation of Pairing Proto-RNA Nucleotides and Their Supramolecular Assemblies, *Life*, 2023, **13**, 2200.

43) Q. Li, J. Zhao, L. Liu, S. Jonchhe, F. J. Rizzuto, S. Mandal, H. He, S. Wei, H. F. Sleiman, H. Mao and C. Mao, A poly(thymine)–melamine duplex for the assembly of DNA nanomaterials, *Nat. Mater.*, 2020, **19**, 1012–1018.

44) G. E. Plum, Y. W. Park, S. F. Singleton, P. B. Dervan and K. J. Breslauer, Thermodynamic characterization of the stability and the melting behavior of a DNA triplex: a spectroscopic and calorimetric study., *Proc. Natl. Acad. Sci.*, 1990, **87**, 9436–9440.

45) A. Ono, P. O. P. Ts'o and L. S. Kan, Triplex formation of an oligonucleotide containing 2'-O-methylpseudoisocytidine with a DNA duplex at neutral pH, *J. Org. Chem.*, 1992, **57**, 3225–3230.

46) M. Egholm, L. Christensen, K. L. Deuholm, O. Buchardt, J. Coull and P. E. Nielsen, Efficient pH-independent sequence-specific DNA binding by pseudoisocytosine-containing bis-PNA, *Nucleic Acids Res.*, 1995, **23**, 217–222.

47) A. B. Eldrup, B. B. Nielsen, G. Haaima, H. Rasmussen, J. S. Kastrup, C. Christensen and P. E. Nielsen, 1,8-Naphthyridin-2(1H)-ones − Novel Bicyclic and Tricyclic Analogues of Thymine in Peptide Nucleic Acids (PNAs), *Eur. J. Org. Chem.*, 2001, **2001**, 1781–1790.

48) G. Devi, Z. Yuan, Y. Lu, Y. Zhao and G. Chen, Incorporation of thio-pseudoisocytosine into triplex-forming peptide nucleic acids for enhanced recognition of RNA duplexes, *Nucleic Acids Res.*, 2014, **42**, 4008–4018.

49) Q. Ma, D. Lee, Y. Q. Tan, G. Wong and Z. Gao, Synthetic genetic polymers: advances and applications, *Polym. Chem.*, 2016, **7**, 5199–5216.

50) C. R. Geyer, T. R. Battersby and S. A. Benner, Nucleobase Pairing in Expanded Watson-Crick-like Genetic Information Systems, *Structure*, 2003, **11**, 1485–1498.

51) S. Hildbrand and C. Leumann, Enhancing DNA Triple Helix Stability at Neutral pH by the Use of Oligonucleotides Containing a More Basic Deoxycytidine Analog, *Angew. Chem. Int. Ed. Engl.*, 1996, **35**, 1968–1970.

52) D. A. Rusling, Four base recognition by triplex-forming oligonucleotides at physiological pH, *Nucleic Acids Res.*, 2005, **33**, 3025–3032.

53) R. T. Ranasinghe, D. A. Rusling, V. E. C. Powers, K. R. Fox and T. Brown, Recognition of CG inversions in DNA triple helices by methylated 3H-pyrrolo[2,3-d]pyrimidin-2(7H)-one nucleoside analogues, *Chem. Commun.*, 2005, 2555.

54) P. Gupta, T. Zengeya and E. Rozners, Triple helical recognition of pyrimidine inversions in polypurine tracts of RNA by nucleobase-modified PNA, *Chem. Commun.*, 2011, **47**, 11125.

55) S. Hildbrand, A. Blaser, S. P. Parel and C. J. Leumann, 5-Substituted 2-Aminopyridine *C* -Nucleosides as Protonated Cytidine Equivalents: Increasing Efficiency and Selectivity in DNA Triple-Helix Formation, *J. Am. Chem. Soc.*, 1997, **119**, 5499–5511.

56) J.-A. Ortega, J. R. Blas, M. Orozco, A. Grandas, E. Pedroso and J. Robles, Binding Affinities of Oligonucleotides and PNAs Containing Phenoxazine and G-Clamp Cytosine Analogues Are Unusually Sequence-Dependent, *Org. Lett.*, 2007, **9**, 4503–4506.

57) Editorial, Catalyzing collaboration, *Nat. Chem. Biol.*, 2007, **3**, 239–239.

58) S. Kuchin, Covering All the Bases in Genetics: Simple Shorthands and Diagrams for Teaching Base Pairing to Biology Undergraduates, *J. Microbiol. Biol. Educ.*, 2011, **12**, 64–66.

59) C. Roberts, R. Bandaru and C. Switzer, Theoretical and Experimental Study of Isoguanine and Isocytosine: Base Pairing in an Expanded Genetic System, *J. Am. Chem. Soc.*, 1997, **119**, 4640–4649.

60) B. A. Schweitzer and E. T. Kool, Hydrophobic, Non-Hydrogen-Bonding Bases and Base Pairs in DNA, *J. Am. Chem. Soc.*, 1995, **117**, 1863–1872.

61) F. Li, P. S. Pallan, M. A. Maier, K. G. Rajeev, S. L. Mathieu, C. Kreutz, Y. Fan, J. Sanghvi, R. Micura, E. Rozners, M. Manoharan and M. Egli, Crystal structure, stability and in vitro RNAi activity of oligoribonucleotides containing the ribo-difluorotoluyl nucleotide: insights into substrate requirements by the human RISC Ago2 enzyme, *Nucleic Acids Res.*, 2007, **35**, 6424–6438.

62) V. Kumar and E. Rozners, Fluorobenzene Nucleobase Analogues for Triplex-Forming Peptide Nucleic Acids, *ChemBioChem*, 2022, **23**, e202100560.

63) A. D. Boese, Density Functional Theory and Hydrogen Bonds: Are We There Yet?, *ChemPhysChem*, 2015, **16**, 978–985.

64) H. Kwon, A. Kumar, M. Del Ben and B. M. Wong, Electron/Hole Mobilities of Periodic DNA and Nucleobase Structures from Large-Scale DFT Calculations, *J. Phys. Chem. B*, 2023, **127**, 5755–5763.

65) P. Mazumdar, A. Kashyap and D. Choudhury, Investigation of hydrogen bonding in small nucleobases using DFT, AIM, NCI and NBO technique, *Comput. Theor. Chem.*, 2023, **1226**, 114188.

66) I. V. Kutyavin, R. L. Rhinehart, E. A. Lukhtanov, V. V. Gorn, R. B. Meyer and H. B. Gamper, Oligonucleotides Containing 2-Aminoadenine and 2-Thiothymine Act as Selectively Binding Complementary Agents, *Biochemistry*, 1996, **35**, 11170–11176.

67) K. G. Devine and S. Jheeta, De Novo Nucleic Acids: A Review of Synthetic Alternatives to DNA and RNA That Could Act as Bio-Information Storage Molecules, *Life*, 2020, **10**, 346.

68) G. Haaima, Increased DNA binding and sequence discrimination of PNA oligomers containing 2,6-diaminopurine, *Nucleic Acids Res.*, 1997, **25**, 4639–4643.

69) J. Lohse, O. Dahl and P. E. Nielsen, Double duplex invasion by peptide nucleic acid: A general principle for sequence-specific targeting of double-stranded DNA, *Proc. Natl. Acad. Sci.*, 1999, **96**, 11804–11808.

70) Y. Zhou, X. Xu, Y. Wei, Y. Cheng, Y. Guo, I. Khudyakov, F. Liu, P. He, Z. Song, Z. Li, Y. Gao, E. L. Ang, H. Zhao, Y. Zhang and S. Zhao, A widespread pathway for substitution of adenine by diaminopurine in phage genomes, *Science*, 2021, **372**, 512–516.

71) D. Sleiman, P. S. Garcia, M. Lagune, J. Loc'h, A. Haouz, N. Taib, P. Röthlisberger, S. Gribaldo, P. Marlière and P. A. Kaminski, A third purine biosynthetic pathway encoded by aminoadenine-based viral DNA genomes, *Science*, 2021, **372**, 516–520.

72) V. Pezo, F. Jaziri, P.-Y. Bourguignon, D. Louis, D. Jacobs-Sera, J. Rozenski, S. Pochet, P. Herdewijn, G. F. Hatfull, P.-A. Kaminski and P. Marliere, Noncanonical DNA polymerization by aminoadenine-based siphoviruses, *Science*, 2021, **372**, 520–524.

73) T. Zengeya, P. Gupta and E. Rozners, Triple-Helical Recognition of RNA Using 2-Aminopyridine-Modified PNA at Physiologically Relevant Conditions, *Angew. Chem. Int. Ed.*, 2012, **51**, 12593–12596.

74) I. Kumpina, N. Brodyagin, J. A. MacKay, S. D. Kennedy, M. Katkevics and E. Rozners, Synthesis and RNA-Binding Properties of Extended Nucleobases for Triplex-Forming Peptide Nucleic Acids, *J. Org. Chem.*, 2019, **84**, 13276–13298.

75) P. Champagne, J. Desroches and J.-F. Paquin, Organic Fluorine as a Hydrogen-Bond Acceptor: Recent Examples and Applications, *Synthesis*, 2014, **47**, 306–322.

Title: **Beyond DAD: Proposing a one-letter code for nucleobase-mediated molecular recognition**
Authors: Aiden J. Ward and Benjamin E. Partridge

Data Availability Statement

No primary research results, software or code have been included and no new data were generated or analyzed as part of this submission.