Importance of Model Size in Quantum Mechanical Studies of DNA Intercalation

Drew P. Harding,[†] Laura J. Kingsley,[§] Glen Spraggon,[§] and Steven E. Wheeler[†*]

[†]*Department of Chemistry, Texas A&M University, College Station, TX 77842 and Center for Computational Quantum Chemistry, Department of Chemistry, University of Georgia, Athens, GA 30602*
E-mail: swheele2@uga.edu
[§]*Genomics Institute of the Novartis Research Foundation, San Diego, CA*

Abstract

There is currently a dearth of effective computational tools to design nucleobase-targeting small molecules and molecular mechanics force-fields for nucleobases lag behind their protein-focused counterparts. While quantum chemical methods can provide reliable interaction energies for small molecule-nucleobase interactions, these come at a steep computational cost. As a first step toward refining available tools for predicting small molecule-nucleobase interactions, we assessed the convergence of DFT-computed interaction energies with increasing binding site model size. We find that while accurate intercalator interaction energies can be derived from binding site models featuring only the flanking nucleotides for uncharged intercalators that bind parallel to the DNA base pairs, errors remain significant even when including distant nucleotides for intercalators that are charged, exhibit groove-binding tails that engage in non-covalent interactions with distant nucleotides, or that bind perpendicular to the DNA base pairs. Consequently, binding site models that include at least three adjacent nucleotides are required to consistently predict converged binding energies. The computationally inexpensive HF-3c method is shown to provide reliable interaction energies and can be routinely applied to such large models.

I. Introduction

Some of the various biological roles of DNA[1] can be modulated by the non-covalent binding of small molecules. A key group of such small molecules are DNA intercalators, which insert between adjacent base pairs leading to partial unwinding and lengthening of the structure of DNA (see Figure 1). These deformed structures can hinder key biological functions, which can be exploited for therapeutic purposes. For instance, intercalators such as doxorubicin and proflavine have been used to treat acute leukemia and the parasite trypanosomiasis, respectively.[2-3] As the biological importance of intercalators has become more widely appreciated and their use as clinical drugs realized, there has been increased interest in the design of intercalator-based therapeutics.[4]

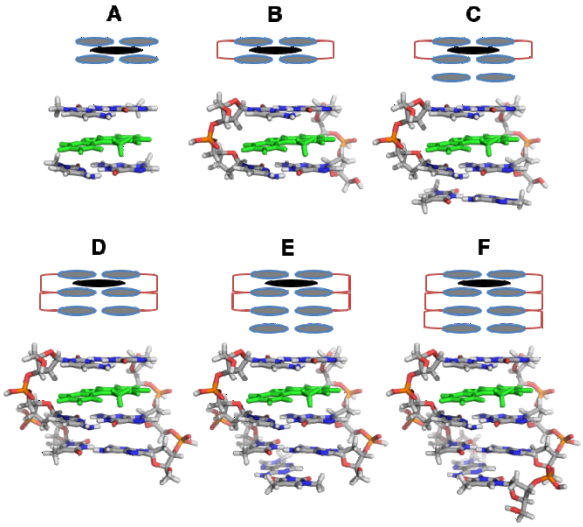

Figure 1. DNA intercalator binding site Models A-F (the intercalator is shown in green).

Despite advances in computational tools for designing small molecules that bind proteins, analogous tools for DNA-targeting small molecules are less well developed. Reasons for this are numerous both from a historical and physiological perspective. Historically, drug design has focused on proteins as drug targets, leading to the development of protein-centric computational tools and structures. From a purely physical standpoint, the flexible structure, helical framework, and high charge density of DNA coupled with numerous binding sites that are often exposed to solvent make specific DNA intercalation a difficult computational problem.[5] The unique features of DNA in comparison to proteins or peptides make it difficult to apply existing computational drug discovery tools to DNA.

Docking is a key example of a computational tool that while widely used in protein based computational drug discovery is difficult to apply in the context of DNA. Given that most docking

protocols utilize a rigid model approach, a binding pocket must already be present. This requires either a crystal structure of intercalated DNA or the creation of an artificial binding pocket.[6-7] Moreover, the scoring functions used to rank binding poses still lack sufficient accuracy to provide reliable predictions for DNA intercalation.[6] At the same time, classical molecular mechanics (MM) force fields have only recently achieved sufficient accuracy to reliably reproduce stable canonical DNA structures over significant simulation times.[8] Since these force fields were parameterized for canonical DNA structures, it is unlikely that they will adequately account for the structural changes that accompany intercalation, let alone reliably predict the strength and specificity of intercalator binding. One route to improved docking scoring functions and MM potentials for DNA intercalation is through parameterization based on reliable quantum mechanical (QM) computations. However, this requires that a sufficiently reliable yet computationally tractable QM method be identified.[9]

There have been numerous QM studies of intercalators.[10-28] For example, In 2011, Pulay *et al.*[15] reported an analysis of the intercalation of ethidium into a UA/AU step of RNA using second order Møller-Plesset perturbation theory (MP2) with the 6-31G(*d*,*p*) basis set in an effort to quantify the importance of pairwise interactions with individual RNA components (nucleobases, sugar-phosphate backbones, *etc*). To this end, they considered truncated models containing the adjacent base pairs with and without the sugar-phosphate backbone. While they were able to devise fragmentation schemes that reproduced the interaction energy of the intact system, they noted that the success of such schemes will vary on a case-by-case basis.[15] That same year, Sherrill *et al.*[16] investigated protonated proflavine intercalated into a CGA trinucleotide using functional-group based symmetry-adapted perturbation theory (F-SAPT). They found that including the backbone in such computations is essential to compute reliable interaction energies, because the interaction of the intercalator with the backbone contributed a third of the total interaction energy. In 2016, Kokoschka *et al*.[29] compared QM intercalator-nucleobase interaction energies to the change in electrochemical impedance spectroscopy measurements of DNA from intercalation. The QM computations included SCS(MI)-MP2/cc-pVTZ interaction energies evaluated at PBE-D3/cc-pVTZ optimized structures of proflavine, ellipticine, and 1-pyrenemethylamine interacting with a single base pair or a single nucleobase as models of double and single stranded DNA. They concluded that these models were enough to capture qualitative trends observed in their experimental results. While these and other studies[10-28] have provided key insights, there exist no

guidelines for the number of bases and backbones that must be included to compute reliable intercalator interaction energies. Moreover, there have been no reported attempts to assess the performance of different DFT methods specifically for intercalators, leaving one to rely on more general benchmarks of stacking interactions to guide the selection of DFT functional for studying DNA intercalation.

Herein, we show that errors due to the model size show unexpectedly slow convergence for certain classes of intercalators, requiring large binding site models in order to reliably predict converged intercalator binding energies. We also show that the minimal basis method HF-3c provides reliable binding energies yet is sufficiently computationally inexpensive to be routinely applied to large binding site models.

II. Theoretical Methods

Structures of ten structurally distinct intercalators bound to 12 binding sites were selected from the PDB: 3FT6,[30] 1K9G,[31] 1Z3F,[32] 4L5K,[33] 465D,[34] 151D,[35] 1DA9,[36] 386D,[37] 110D,[36] 1VTH,[38] 1DL8,[39] and 1NAB[40] (see Figure 2). In nine of these structures, intercalation occurs between GC bases, which is reflective of the PDB in that the binding sites are heavily biased to intercalation at GC sites.[41] Further details about the intercalation sites and the binding stoichiometry of the selected systems can be found in SI Table S1.

Atoms not covalently bound to either DNA or the intercalator (*e.g.* solvent, counterions, *etc*.) were removed and all phosphate groups protonated.[42-44] For each intercalated DNA structure, six models were constructed (Model A-F in Figure 1). The largest of these consists of intercalated DNA truncated to the four base pairs surrounding the intercalators (see Model F in Figure 1). It should be noted that all intercalators considered were bound between the two terminal bases of the helical strand, which is reflective of the binding position in the crystal structures. Hydrogen atoms were added to satiate all open valences and the positions of all hydrogen atoms were optimized at the B97-D/6-31G(d) level of theory[45-47] with the coordinates of the heavy atoms frozen. Some intercalators included ionizable nitrogen atoms. We considered protonated versions of these intercalators, which are denoted with a + (*e.g.* $\mathbf{3}^{+}$, see Figure 3).

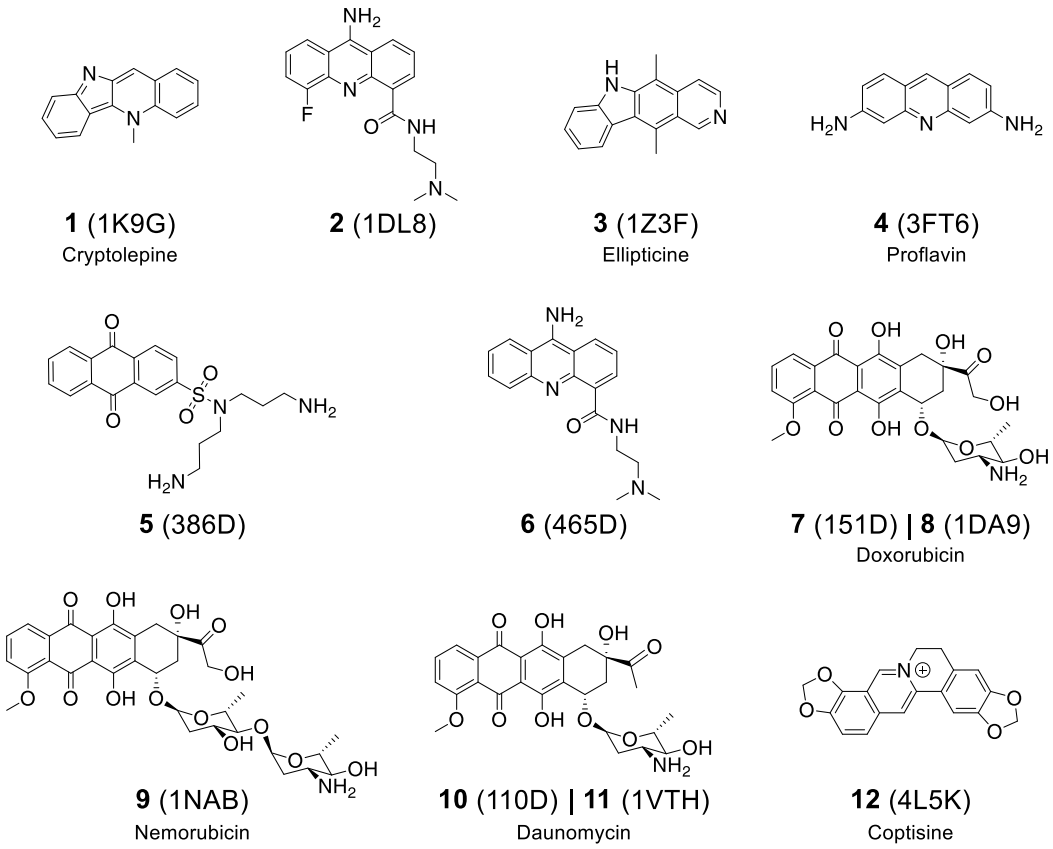

Figure 2. Intercalators studied along with corresponding PDB codes. For intercalators (**7**, **8**) and (**10**, **11**), the same intercalator is bound to two different binding pockets. Intercalators **1**-**3** were used to assess DFT methods, whereas all 12 were considered for assessing the convergence of binding energies with respect to model size.

Interaction energies, defined as the difference in energy between the intercalated DNA and the isolated DNA and intercalator at the hydrogen-optimized crystal-structure geometry, were evaluated using rigorous *ab initio* methods, DFT methods, and the parameterized methods HF-3c and PBEh-3c. First, reliable interaction energies were determined within the focal point approach of Allen and co-workers[48-50] for four intercalators using the two smallest binding site models. In the focal point approach, dual extrapolations with respect to one-particle basis set and electron correlation were carried out using RI-MP2,[51] DLPNO-CCSD,[52-55] and DLPNO-CCSD(T)[52-55] with the Dunning correlation-consistent basis sets up to cc-pVQZ.[56] In particular, Hartree-Fock energies were extrapolated based on cc-pVXZ (X = D, T, Q) energies using an exponential functional form[57] while correlation energies were extrapolated separately from cc-pVXZ (X = T, Q) data following Helgaker *et al.*[58] Sponer and co-workers[59] recently demonstrated the reliability of extrapolated DLPNO-CC interaction energies on model stacked nucleotides. Incremental focal point tables are included as SI Table S2 and suggest that these interaction energies are converged

to within 2.5 kcal mol^{-1} of the *ab initio* limit. DFT interaction energies were computed using the B3LYP-D3,[60-61] B97-D,[45-46] M06-2X,[62] and ωB97X-D[63] functionals in combination with the 6-31G(d),[47] 6-31+G(d),[47] def2-TZVP,[64] and cc-pVTZ[56] basis sets. The B3LYP-D3 data include three-body dispersion corrections and used the Becke-Johnson damping function.[65] Interaction energies were also computed using HF-3c and PBEh-3c.[66-67] The electrostatic component of these interaction energies was estimated by first computing the electrostatic potential (ESP) due to the DNA at the positions of the intercalator atoms. The total electrostatic interaction energy was estimated as the sum of products of NPA atomic charges[68-69] for the isolated intercalator and the ESP values at the corresponding positions. All DFT computations carried out with Gaussian09[70] while Orca 4.0[71] was used to compute RI-MP2, DLPNO-CCSD, DLPNO-CCSD(T), HF-3c, and PBEh-3c energies. The DLPNO computations utilized default PNO cutoffs.

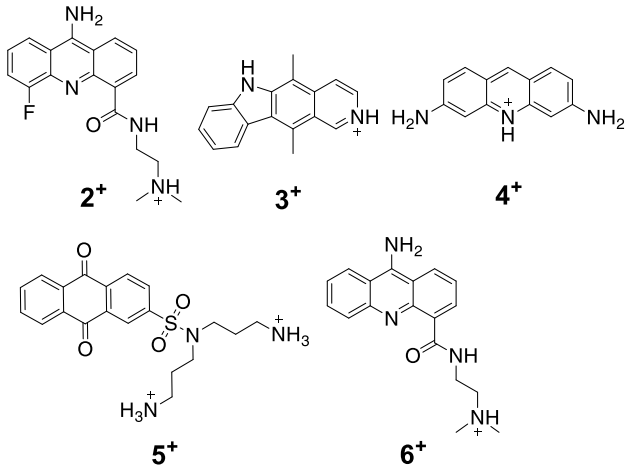

Figure 3. Protonated forms of intercalators **2-6**.

III. Results and Discussion

Ten intercalators (see Figure 2) were selected to represent typical intercalators found on the PDB. They range from planar anthracene-based structures (*e.g.* proflavine, **4**) to tetracyclic structures appended with flexible groove-binding tails (*e.g.* doxorubicin, **7**). Representative binding modes are depicted in Figure 4. The intercalated DNA crystal structures were truncated to six distinct models ranging from two to four base pairs both with and without the sugar-phosphate backbone (Models A-F, Figure 1). The impact of both level of theory and model size on computed interaction energies were quantified.

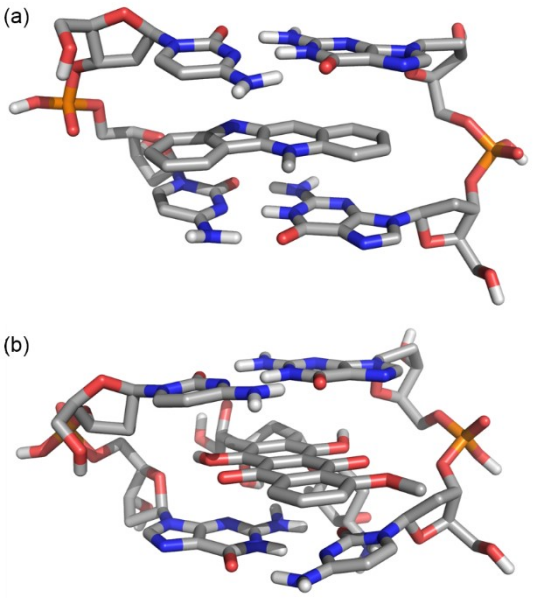

Figure 4. Representative parallel and perpendicular binding modes for (a) **1** and (b) **7**, respectively.

A. Performance of DFT Methods

We first assess the performance of popular DFT methods in predicting intercalator interaction energies. Given the absence of experimental binding energies for the intercalated systems for which reliable crystal structures are available, we used DLPNO-CCSD(T) interaction energies extrapolated to the complete basis set (CBS) limit for intercalators **1**-**3**, as well as $\mathbf{3}^{+}$, using the two smallest binding site models (Models A and B) to assess the performance of different DFT methods. As shown below, for these four systems Model B captures 95% of the interaction energy of the most complete model (Model F). The computed reference interaction energies are listed in Table 1.

Sixteen combinations of four DFT functions paired with four popular basis sets were explored. Figure 5 shows the percent error of the DFT-predicted interaction energies compared to the *ab initio* data for **1**-**3** and $\mathbf{3}^{+}$ using Model B; a table of the values can be found in SI Table S3. Results for Model A are similar (see SI Table S4). There is considerable spread in errors in the DFT interaction energies, which can exceed 30%. This should serve as a clear warning that the poor choice of either functional or basis set for such systems can lead to significant errors in computed interaction energies. However, it should be noted that even these errors are considerably smaller than those using MP2 (see SI Table S5), which is widely used in the literature.[11,15,21-27] Fortunately, for many functional/basis set combinations the errors are considerably smaller, sometimes below 1%. Overall, several trends emerge from these data, if M06-2X is excluded.

First, errors for a given functional generally decrease going from 6-31G(d) to 6-31+G(d), then cc-pVTZ, and finally to def2-TZVP. The basis set dependence of the M06-2X predictions are almost exactly the opposite—errors are maximized with def2-TZVP and minimized with the smaller Pople-style basis sets. While not shown in Figure 5, we also considered B3LYP without an empirical dispersion correction. The resulting interaction energies for these dispersion-dominated systems exhibited errors as large as 140%. Simply put, B3LYP (and other functionals that do not capture dispersion interactions) should not be used to study DNA intercalation. In contrast to this, B3LYP-D3/def2-TZVP provides the most reliable interaction energies for these systems, with errors consistently below 5% (see Table 1). For instance, the largest error (1.3 kcal mol^{-1}) occurs for $\mathbf{3}^+$ with Model B; errors are all considerably less than 1 kcal mol^{-1} for the other systems considered. Other levels of theory that perform nearly as well include B97-D/def2-TZVP and M06-2X/6-31+G(d).

Table 1. B3LYP-D3(BJ)/def2-TZVP interaction energies (kcal mol^{-1}) for intercalators **1-12** using models A-F. DLPNO-CCSD(T)/CBS interaction energies are provided for selected systems in parentheses. Whether the intercalator binds parallel (par) or perpendicular (perp) to the base pair axes is also indicated.

	Binding Mode	A	B	C	D	E	F
1	par	-41.1 (-42.0)	-45.2 (-45.4)	-46.0	-45.9	-45.6	-45.9
2	par	-39.9 (-39.5)	-45.7 (-45.3)	-46.4	-46.2	-46.8	-46.6
$\mathbf{2}^+$	par	-72.8	-86.5	-90.4	-92.4	-93.8	-95.4
3	par	-34.4 (-34.6)	-40.2 (-39.4)	-41.2	-41.4	-41.4	-41.2
$\mathbf{3}^+$	par	-47.9 (-47.7)	-63.8 (-62.5)	-64.2	-65.6	-66.8	-67.9
4	par	-36.1	-45.5	-46.2	-46.7	-46.6	-46.7
$\mathbf{4}^+$	par	-51.8	-69.7	-71.5	-73.0	-73.6	-75.1
5	par	-41.4	-48.5	-49.0	-49.3	-49.6	-49.6
$\mathbf{5}^+$	par	-55.1	-71.9	-75.3	-79.1	-80.5	-83.3
6	par	-40.6	-46.5	-46.6	-47.1	-47.6	-47.8
$\mathbf{6}^+$	par	-77.3	-87.4	-90.9	-93.1	-94.4	-95.1
7	perp	-42.2	-61.0	-67.6	-73.9	-74.2	-76.0
8	perp	-35.8	-46.1	-50.9	-54.6	-55.6	-56.2
9	perp	-52.5	-63.8	-65.0	-69.2	-73.1	-73.3
10	perp	-51.4	-58.1	-60.2	-62.6	-63.0	-63.1
11	perp	-50.8	-40.6	-45.6	-48.7	-49.4	-53.3
12	perp	-38.5	-45.6	-48.3	-50.6	-51.4	-53.6

We also considered two parameterized minimal-basis set methods designed to provide reliable interaction energies at a much lower computational cost, HF-3c and PBEh-3c.[66-67] Data for HF-3c are provided in Table 2 (See SI Table S6 for PBEh-3c results). Percent errors in interaction energies, relative to the DLPNO-CCSD(T) data, are plotted in Figure 5 for these four systems using Model B. While errors for PBEh-3c range from 7 to 17%, HF-3c provides

interaction energy errors for **1-3** and $\mathbf{3}^{+}$ consistently below 6%. This is well below the errors observed for many of the DFT/basis set combinations.

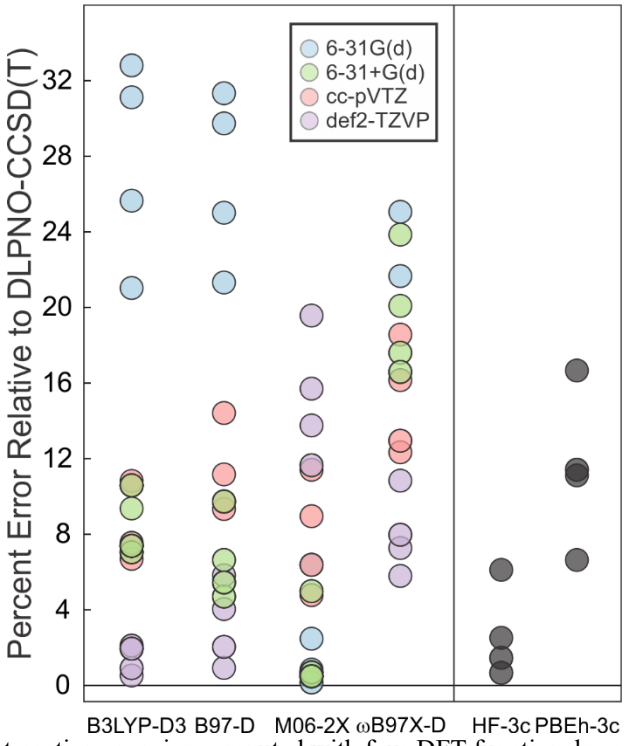

Figure 5. Percent errors in interaction energies computed with four DFT functionals paired with four popular basis sets as well as HF-3c and PBEh-3c compared to extrapolated DLPNO-CCSD(T) values for intercalators **1-3** and $\mathbf{3}^{+}$ using Model B. Data for Model A are available in SI.

Table 2. HF-3c interaction energies (kcal mol^{-1}) using Models A-F.

	A	B	C	D	E	F
1	-40.3	-44.3	-45.2	-45.3	-45.3	-45.7
2	-35.7	-42.5	-43.8	-43.4	-44.0	-43.6
$\mathbf{2}^{+}$	-70.8	-85.3	-90.3	-92.9	-94.2	-96.0
3	-32.4	-39.1	-40.2	-40.3	-40.4	-40.3
$\mathbf{3}^{+}$	-44.5	-61.6	-62.0	-63.5	-64.9	-65.9
4	-34.8	-44.6	-45.4	-45.7	-45.7	-45.6
$\mathbf{4}^{+}$	-49.0	-67.3	-69.4	-71.0	-71.6	-73.0
5	-38.8	-46.2	-47.4	-47.7	-48.0	-48.3
$\mathbf{5}^{+}$	-49.3	-65.2	-69.5	-73.1	-74.5	-77.5
6	-37.0	-44.0	-44.4	-44.9	-45.4	-45.4
$\mathbf{6}^{+}$	-76.4	-87.7	-92.5	-94.8	-96.1	-96.7
7	-37.5	-56.1	-62.8	-68.8	-69.1	-70.8
8	-32.5	-42.3	-46.6	-50.6	-51.6	-52.7
9	-47.7	-57.1	-58.2	-62.4	-66.4	-66.5
10	-45.6	-50.7	-53.6	-56.4	-56.8	-57.2
11	-44.9	-30.4	-34.6	-37.2	-38.1	-42.0
12	-34.6	-42.0	-45.0	-47.9	-48.7	-50.6

B. Convergence with Model Size

Computational expediency necessitates the use of a truncated models (*i.e.* using a small portion of the DNA) in applications of QM methods to DNA intercalation. Unfortunately, there has been no study that has identified the minimal model required to reliably capture interactions between an intercalator and DNA. To address this, we computed the interaction energy of each intercalator using six models of increasing size (Models A-F; see Figure 1). The models, which use coordinates from available crystal structures,[30-40] range from the inclusion of only two base pairs (Model A) to four pairs of tetranucleotides (Model F). Interaction energies using these six models were computed for the 12 non-protonated intercalators for all 16 levels of DFT. Given the small errors observed for B3LYP-D3/def2-TZVP above, we focus on these data below; errors in predicted interaction energies for the 15 other levels of DFT can be found in SI Table S6 (the trends discussed below apply across all DFT methods considered). B3LYP-D3/def2-TZVP computed interaction energies for all intercalators considered based on Models A-F are listed in Table 1. Overall, the gas-phase interaction energies of the non-protonated intercalators with Model F range from –41.2 kcal mol^{-1} to –76.0 kcal mol^{-1}.

Percent absolute errors in predicted interaction energies, compared to Model F, are plotted in Figure 6a as a function of model size. Excluding the obvious outlier of **11** with Model A, which can be attributed to a fortuitous cancelation of sizeable errors, these interaction energies show monotonic convergence with increasing model size. Notably, errors for the smallest model size (Model A, which features only the adjacent base pairs) are large, ranging from 10 to nearly 50% of the total interaction energy. Inclusion of the sugar-phosphate backbone (*i.e.* Model B) leads to drastically reduced errors, which are below 25% for all 12 systems. For some systems (**1**-**6**), errors using Model B drop by nearly an order of magnitude compared to Model A, to well below 5% (see Figure 6b). This is consistent with previous work[15-16] showing that the backbone contributes significantly to the total interaction energy. However, for five of the systems (**7**-**12**), errors in predicted interaction energies remain above 10% with Model B (corresponding to errors in interaction energies of 5-8 kcal mol^{-1}) and show a very slow convergence with model size (see Figure 6c). Even with Model D, which includes three base pairs and the associated sugar-phosphate backbones, errors exceed 5% for three of these systems (**9**, **11**, and **12**).

Errors for **10** and **11** are instructive because they involve the same intercalator bound to two different sites. While errors for **10** drop below 5% after Model C, **11** shows much slower convergence with respect to the inclusion of distant nucleotides. Indeed, even for Model E the

errors for **11** are in excess of 5%. This difference in behavior can be attributed to the position of the pendant sugar group on this intercalator in the two binding sites. In **11**, this group is engaged in several H-bonding interactions with a distant sugar-phosphate backbone, whereas it interacts with more proximal groups in the case of **10**. The result is that inclusion of four nucleotide pairs plus all sugar-phosphate backbones (*i.e.* Model F) is mandatory to capture the interaction energy in **11**; for **10**, smaller models suffice.

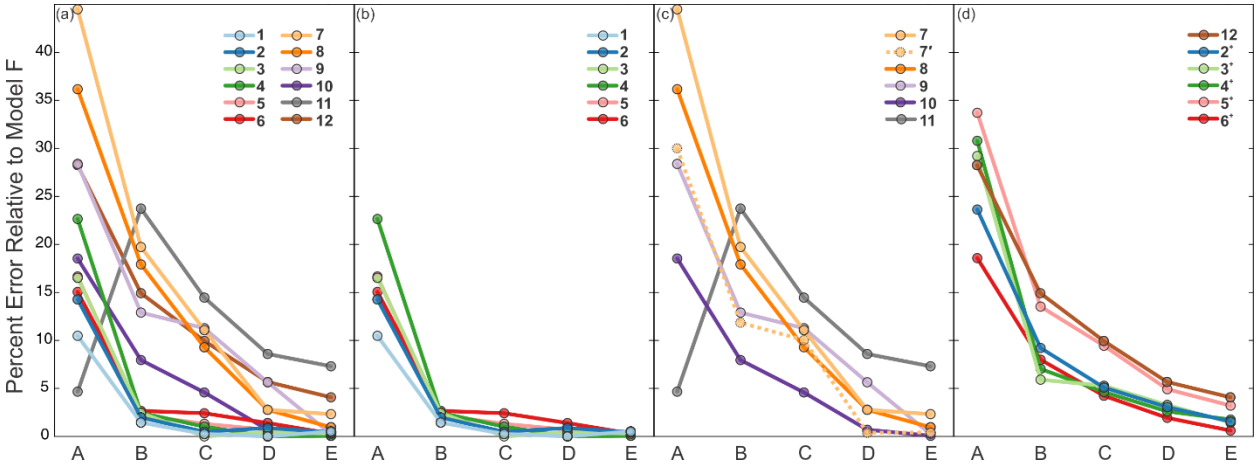

Figure 6. Percent absolute error of Model A-E, relative to Model F, at the B3LYP-D3/def2-TZVP level of theory for (a) all unprotonated intercalators (**1**-**12**), (b) intercalators that exhibit rapid convergence with respect to model size (**1**-**6**), (c) neutral intercalators that show slower convergence with respect to model size (**7**-**11**), and (d) charged intercalators (**12**, **2**$^+$-**6**$^+$).

From a broader perspective, there are no obvious structural features delineating intercalators that exhibit slow vs fast convergence with model size. For instance, errors for **1**-**6** are all below 5% for Model B, but these intercalators have drastically different molecular structures. In particular, the intercalators in **2**, **5**, and **6** include groove-binding tails, yet the interaction energy is well-converged even with Model B. The main common feature of these six systems is that they all bind nearly parallel to the base pairs (*e.g.* Figure 4a). Moreover, for the intercalators with groove-binding tails, these tails mainly interact with the adjacent nucleobases, rendering Model B sufficient for capturing all direct non-covalent interactions.

Systems **7**-**11**, on the other hand, all include groove-binding tails and all bind nearly perpendicular to the base pairs (*e.g.* Figure 4b). To differentiate between these two effects, we considered a truncated version of **7** (**7′**), in which the sugar group was removed and replaced with a hydrogen. While removal of this pendant group results in some decrease in percent error (see Figure 6c), compared to **7**, the convergence with model size for **7′** still resembles the other

perpendicular intercalators (**8**-**11**). This, combined with the comparison of **10** and **11**, suggests that the slow convergence with respect to model size exhibited by **7**-**11** is due to both the groove-binding tails and the perpendicular binding modes. The former effect is most severe when the groove-binding tail interacts with distant nucleotides.

To further characterize the source of this slow convergence for selected intercalators, we approximated the electrostatic contribution to the total interaction energies (see Computational Methods for details). The absolute differences in the electrostatic contribution ($|\Delta E_{elec}|$), compared to Model F, are plotted in Figure 7. Errors in the electrostatic component mirror those observed for the total interaction energy—they converge rapidly for **1**-**6** (see Figure 7a) but much more slowly for **7**-**11** (see Figure 7b). The large $|\Delta E_{elec}|$ values for Model A are unsurprising given Sherrill's demonstration[16] of the large electrostatic interaction between intercalators and the backbone, which is absent in Model A. For **1**-**6**, these errors are mostly eliminated by Model B. This rapid convergence of $|\Delta E_{elec}|$ for the parallel-binding intercalators can be attributed to the fact that they are buried between flanking nucleobases, so through-space electrostatic interactions with distant nucleotides will be screened by the intervening electron density. For **7**-**11**, on the other hand, significant $|\Delta E_{elec}|$ values remain even for the larger models. This can be explained by the presence of unimpeded through-space electrostatic interactions with distant backbone and nucleobase atoms for the systems that bind perpendicular to the base pairs. Thus, the slow convergence of QM interaction energies for the perpendicular intercalators is partly an artifact of the use of gas-phase computations. Consequently, solution-phase computations that include counterions for systems such as **7**-**11** could converge more quickly with respect to model size, because the intervening solvent and ions will provide some dielectric screening.

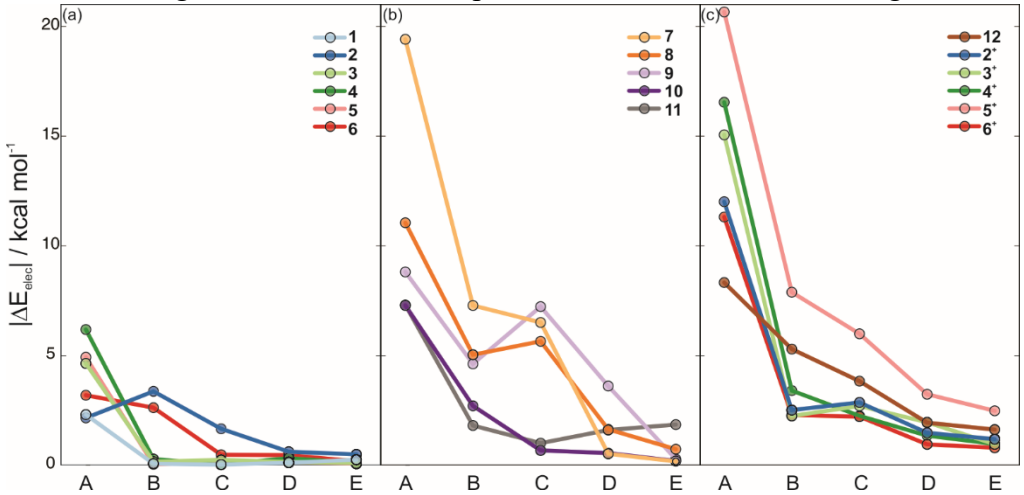

Figure 7. Absolute errors in the electrostatic component of the total interaction energy, relative to Model F, for a) **1**-**6,** b) **7**-**11**, and c) **12** and **2**$^{+}$-**6**$^{+}$.

A key potential use of QM for studying intercalators is ranking intercalators by their interaction strength. To this end, we assessed the reliability of rankings from each model size compared to Model F (see Figure 8). Rankings from Model A are qualitatively incorrect. For instance, using Model F, **1** is the 11th weakest binder, whereas Model A predicts it is ranked 6th. Indeed, only **10** and **12** are ranked correctly using Model A. While the rank order from Model B is considerably better, mirroring the drastic drop in percent error in interaction energies discussed above, still only **3** and **12** are ranked correctly. By Model C, the ranking is qualitatively correct, except for one outlier (**11**) and the ranking is correct apart from small swaps for Models D and E. As such, Model C should be considered a minimum model for reliably ranking intercalator binding, with Model D providing more robust results.

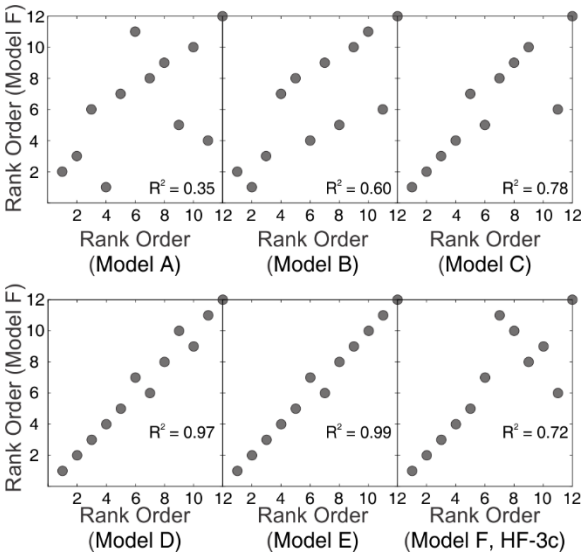

Figure 8. Rank-order correlation for intercalators **1**-**12** based on Models A-E using B3LYP-D3/def2-TZVP and Model F using HF-3c compared to Model F using B3LYP-D3/def2-TZVP.

Charged vs Neutral Intercalators

Of the intercalators considered, only **12** carries a permanent positive charge. This system, which lacks a groove-binding tail yet binds perpendicularly, exhibits similarly slow convergence with respect to model size as seen for intercalators **7**-**11**. To probe the impact of intercalator charge, we considered protonated versions of **2**-**6** (see Figure 3). Interaction energies for **2**$^+$-**6**$^+$ with Models A-F were computed with all 16 DFT methods. B3LYP-D3/def2TZVP results are listed in Table 1 and plotted in Figure 6d; the remaining results can be found in the SI. As expected, there is a sharp increase in interaction energies upon protonation of **2**-**6** across all models (see Table 1). Clearly, the protonation state will have a notable effect and intercalation energies computed using the

incorrect protonation state will be qualitatively incorrect. This increase in interaction energy is accompanied by an increased percent error across all models and a much slower convergence, as seen for **12** (see Figure 6d).

As done for the neutral intercalators, we also estimated the electrostatic component of the total interaction energies for **12** and $\mathbf{2}^{+}$-$\mathbf{6}^{+}$. $|\Delta E_{elec}|$ is considerably larger for the charged vs uncharged intercalators. This is most severe for Model A, which lacks the highly polar sugar-phosphate backbone. As seen for intercalators **7**-**11**, $|\Delta E_{elec}|$ declines for Model B, but remains non-negligible for the larger models. Thus, the large errors for the charged intercalators can also be attributed to long-range electrostatic interactions with distant nucleotides.

Performance of HF-3c

Given that large binding site models are required to provide converged intercalator binding energies, combined with the accuracy of HF-3c demonstrated for selected intercalators using Models A and B, we assessed the performance of this computationally inexpensive method for all models sizes compared to B3LYP-D3/def2-TZVP data. First, HF-3c interaction energies (Table 2) show the same convergence with respect to model size as B3LYP-D3 and the other DFT methods. Considering all 12 systems with all six models, there is a strong correlation between the HF-3c results and those from B3LYP-D3 ($R^2 = 0.97$, see Figure 9), with a mean absolute deviation (MAD) and mean sign deviation (MSD) of 3.4 and –3.2 kcal mol^{-1}, respectively. That is, HF-3c provides interaction energies for these systems that are systematically slightly weaker than those from B3LYP-D3/def2-TZVP, but at a significantly reduced computational cost. While differences between HF-3c and B3LYP-D3 are as large as 11 kcal mol^{-1} for some systems, for other intercalators the differences are consistently below 1 kcal mol^{-1}. Overall, the largest deviations between the HF-3c and B3LYP-D3 data occur for the perpendicular binding intercalators (**7**-**12**), so potentially stems from the underestimation of long-range electrostatic interactions by HF-3c.

With regard to ranking intercalator binding, HF-3c provides correct rankings for the five strongest binding intercalators using Model F (compared to B3LYP-D3) and also correctly predicts the most strongly bound intercalator (see Figure 8). Of the six intercalators that are ranked incorrectly, the spread of energies (4 kcal mol^{-1}) is commensurate with the average error of HF-3c, so its failure to rank these correctly is unsurprising. Overall, these data suggest that HF-3c constitutes a practical approach for studying intercalators that will allow for routine computations on large binding site models.

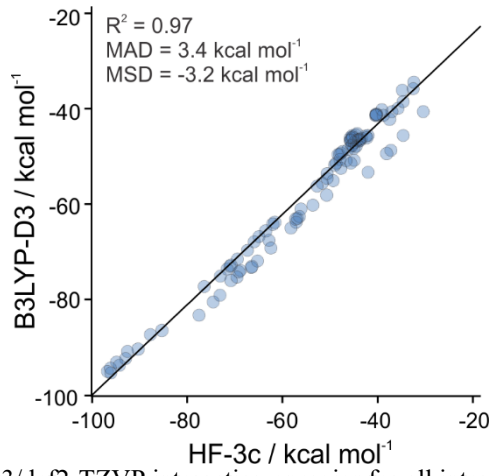

Figure 9. Correlation of B3LYP-D3/def2-TZVP interaction energies for all intercalators using all model sizes with data from HF-3c.

IV. Concluding Remarks

Although quantum mechanical methods can provide reliable interactions for DNA intercalators, this requires the judicious choice of method and model size. Unfortunately, guidelines for such choices were not previously available. We have assessed errors in DFT-predicted intercalator binding energies and the convergence of these errors with respect to binding site size for 12 representative intercalators. Based on comparisons to reliable *ab initio* interaction energies for selected systems and small model sizes, we recommend B3LYP-D3(BJ)/def2-TZVP, B97-D/def2-TZVP, or M06-2X/6-31+G(d) for DFT based studies of DNA intercalation. B3LYP and other functionals that fail to capture dispersion interactions, as well as MP2, provide qualitatively incorrect interaction energies and should be not be applied to intercalated systems. As for model size, recommendations depend on several factors. For neutral intercalators that bind parallel to the base-pairs and lack groove-binding tails that interact with distant nucleotides, errors in predicted interaction energies are already converged upon inclusion of the flanking base-pairs and sugar-phosphate backbone (Model B in Figure 1). However, for intercalators that carry a formal charge, bind perpendicular to the base-pairs, or include groove-binding tails that engage in non-covalent interactions with distant nucleotides, significantly larger models are required. For instance, for some intercalators errors approaching 10% can persist even for Model E. Given the slow convergence of intercalator binding energies in some cases, it is particularly important to identify affordable computational approaches for such systems. To this end, we showed that HF-3c provides interaction energies that are in close agreement with the best tested DFT method across all model sizes, providing a pragmatic method that can be routinely applied to larger binding site

models. Our hope is that these recommendations can help guide future applications of QM methods to intercalation, which in turn can be used to refine computational tools for designing nucleotide targeting small molecules.

Acknowledgments

This work was supported in part by the National Science Foundation (Grant CHE-1807328). Portions of this research were conducted with high performance research computing resources provided by Texas A&M University (http://hprc.tamu.edu) and the Georgia Advanced Computing Resource Center (http://gacrc.uga.edu).

Conflicts of Interest

L.J.K. and G.S. are employees of the Novartis Institute for Biomedical Research and declare no competing financial interests.

Supporting Information Available: Additional computational data, absolute energies, Cartesian coordinates.

References

1. Leung, C. H.; Chan, D. S.; Ma, V. P.; Ma, D. L., DNA-binding small molecules as inhibitors of transcription factors. *Med. Res. Rev.* **2013,** *33*, 823-846.
2. Mross, K.; Massing, U.; Kratz, F., DNA-intercalators — the anthracyclines. In *Drugs Affecting Growth of Tumours. Milestones in Drug Therapy*, Pinedo, H. M.; Smorenburg, C. H., Eds. Birkhäuser Basel: 2006; pp 19-81.
3. Figgitt, D.; Denny, W.; Chavalitshewinkoon, P.; Wilairat, P.; Ralph, R., In vitro study of anticancer acridines as potential antitrypanosomal and antimalarial agents. *Antimicrobial Agents and Chemotherapy* **1992,** *36*, 1644-1647.
4. Mukherjee, A.; Sasikala, W. D., Drug-DNA intercalation: From discovery to the molecular mechanism. In *Advances in Protein Chemistry and Structural Biology*, 2013; Vol. 92, pp 1-62.
5. Rescifina, A.; Zagni, C.; Varrica, M. G.; Pistarà, V.; Corsaro, A., Recent advances in small organic molecules as DNA intercalating agents: Synthesis, activity, and modeling. *Eur. J. Med. Chem.* **2014,** *74*, 95-115.
6. Soni, A.; Khurana, P.; Singh, T.; Jayaram, B., A DNA intercalation methodology for an efficient prediction of ligand binding pose and energetics. *Bioinformatics* **2017,** *33*, 1488-1496.
7. Ricci, C. G.; Netz, P. A., Docking studies on DNA-ligand interactions: building and application of a protocol to identify the binding mode. *J. Chem. Inf. Model.* **2009,** *49*, 1925-1935.

8. Zgarbová, M.; Šponer, J.; Otyepka, M.; Cheatham, T. E.; Galindo-Murillo, R.; Jurečka, P., Refinement of the Sugar-Phosphate Backbone Torsion Beta for AMBER Force Fields Improves the Description of Z- and B-DNA. *Journal of Chemical Theory and Computation* **2015,** *11*, 5723-5736.
9. Mackerell, A. D., Empirical force fields for biological macromolecules: Overview and issues. *Journal of Computational Chemistry* **2004,** *25*, 1584-1604.
10. Jambrec, D.; Haddad, R.; Lauks, A.; Gebala, M.; Schuhmann, W.; Kokoschka, M., DNA Intercalators for Detection of DNA Hybridisation: SCS(MI)-MP2 Calculations and Electrochemical Impedance Spectroscopy. *ChemPlusChem* **2016,** *81*, 604-612.
11. Xu, P.; Wang, J.; Xu, Y.; Chu, H.; Shen, H., Advance in Structural Bioinformatics. 2015; Vol. 827, pp 187-203.
12. Terryn, R. J.; German, H. W.; Kummerer, T. M.; Sinden, R. R.; Baum, J. C.; Novak, M. J., Novel computational study on π -stacking to understand mechanistic interactions of Tryptanthrin analogues with DNA. *Toxicology Mechanisms and Methods* **2014,** *24*, 73-79.
13. Hardebeck, L. K. E.; Johnson, C. A.; Hudson, G. A.; Ren, Y.; Watt, M.; Kirkpatrick, C. C.; Znosko, B. M.; Lewis, M., Predicting DNA-intercalator binding: The development of an arene-arene stacking parameter from SAPT analysis of benzene-substituted benzene complexes. *Journal of Physical Organic Chemistry* **2013,** *26*, 879-884.
14. Hill, G. M.; Moriarity, D. M.; Setzer, W. N., Attenuation of cytotoxic natural product DNA intercalating agents by caffeine. *Scientia Pharmaceutica* **2011,** *79*, 729-747.
15. Langner, K. M.; Janowski, T.; Góra, R. W.; Dziekoński, P.; Sokalski, W. A.; Pulay, P., The Ethidium-UA/AU Intercalation Site: Effect of Model Fragmentation and Backbone Charge State. *Journal of Chemical Theory and Computation* **2011,** *7*, 2600-2609.
16. Hohenstein, E. G.; Parrish, R. M.; Sherrill, C. D.; Turney, J. M.; Schaefer, H. F., Large-scale symmetry-adapted perturbation theory computations via density fitting and Laplace transformation techniques: Investigating the fundamental forces of DNA-intercalator interactions. *Journal of Chemical Physics* **2011,** *135*.
17. Hargis, J. C.; Schaefer, H. F.; Houk, K. N.; Wheeler, S. E., Noncovalent interactions of a benzo[a]pyrene diol epoxide with DNA base pairs: Insight into the formation of adducts of (+)-BaP DE-2 with DNA. *Journal of Physical Chemistry A* **2010**.
18. Grant Hill, J.; Platts, J. A., Local electron correlation descriptions of the intermolecular stacking interactions between aromatic intercalators and nucleic acids. *Chemical Physics Letters* **2009,** *479*, 279-283.
19. Li, S.; Cooper, V. R.; Thonhauser, T.; Lundqvist, B. I.; Langreth, D. C., Stacking interactions and DNA intercalation. *Journal of Physical Chemistry B* **2009,** *113*, 11166-11172.
20. Barone, G.; Guerra, C. F.; Gambino, N.; Silvestri, A.; Lauria, A.; Almerico, A. M.; Bickelhaupt, F. M., Intercalation of daunomycin into stacked dna base pairs. Dft study of an anticancer drug. *Journal of Biomolecular Structure and Dynamics* **2008,** *26*, 115-129.
21. Černý, J.; Hobza, P., Non-covalent interactions in biomacromolecules. *Physical Chemistry Chemical Physics* **2007,** *9*, 5291.
22. Langner, K. M.; Kedzierski, P.; Sokalski, W. A.; Leszczynski, J., Physical nature of ethidium and proflavine interactions with nucleic acid bases in the intercalation plane. *Journal of Physical Chemistry B* **2006,** *110*, 9720-9727.
23. Kubař, T.; Hanus, M.; Ryjáček, F.; Hobza, P., Binding of cationic and neutral phenanthridine intercalators to a DNA oligomer is controlled by dispersion energy: Quantum chemical

calculations and molecular mechanics simulations. *Chemistry - A European Journal* **2005,** *12*, 280-290.
24. Dračínský, M.; Castaño, O., Calculations of interaction energies of ellipticine derivatives with DNA base pairs. *Phys. Chem. Chem. Phys.* **2004,** *6*, 1799-1805.
25. Kumar, A.; Elstner, M.; Suhai, S., SCC-DFTB-D study of intercalating carcinogens: Benzo(a)pyrene and its metabolites complexed with the G-C base pair. *International Journal of Quantum Chemistry* **2003,** *95*, 44-59.
26. Řeha, D.; Kabeláč, M.; Ryjáček, F.; Sponer, J.; Šponer, J. E.; Elstner, M.; Suhai, S.; Hobza, P., Intercalators. 1. Nature of stacking interactions between intercalators (ethidium, daunomycin, ellipticine, and 4′,6-diaminide-2-phenylindole) and DNA base pairs. Ab initio quantum chemical, density functional theory, and empirical potential study. *Journal of the American Chemical Society* **2002,** *124*, 3366-3376.
27. Bondarev, D. A.; Skawinski, W. J.; Venanzi, C. A., Nature of Intercalator Amiloride - Nucelobase Stacking . An Empirical Potential and ab Initio Electron Correlation Study. *J. Phys. Chem. B* **2000,** *104*, 815-822.
28. Medhi, C.; Mitchell, J. B. O.; Price, S. L.; Tabor, A. B., Electrostatic factors in DNA intercalation. *Biopolymers* **1999,** *52*, 84-93.
29. Jambrec, D.; Haddad, R.; Lauks, A.; Gebala, M.; Schuhmann, W.; Kokoschka, M., DNA Intercalators for Detection of DNA Hybridisation: SCS(MI)-MP2 Calculations and Electrochemical Impedance Spectroscopy. *Chempluschem* **2016,** *81*, 604-612.
30. Maehigashi, T.; Persil, O.; Hud, N. V.; Williams, L. D., Crystal Structure of Proflavine in Complex with a DNA hexamer duplex. *To be publsihed*.
31. Lisgarten, J. N.; Coll, M.; Portugal, J.; Wright, C. W.; Aymami, J., The antimalarial and cytotoxic drug cryptolepine intercalates into DNA at cytosine-cytosine sites. *Nat Struct Biol* **2002,** *9*, 57-60.
32. Canals, A.; Purciolas, M.; Aymami, J.; Coll, M., The anticancer agent ellipticine unwinds DNA by intercalative binding in an orientation parallel to base pairs. *Acta Crystallogr D Biol Crystallogr* **2005,** *61*, 1009-1012.
33. Ferraroni, M.; Bazzicalupi, C.; Gratteri, P., Crystal structure of the complex of DNA hexamer d(CGATCG) with Coptisine. *To be published*.
34. Adams, A.; Guss, J. M.; Collyer, C. A.; Denny, W. A.; Wakelin, L. P., Crystal structure of the topoisomerase II poison 9-amino-[N-(2-dimethylamino)ethyl]acridine-4-carboxamide bound to the DNA hexanucleotide d(CGTACG)2. *Biochemistry* **1999,** *38*, 9221-9233.
35. Lipscomb, L. A.; Peek, M. E.; Zhou, F. X.; Bertrand, J. A.; D., V.; L.D., W., Water ring structure at DNA interfaces: hydration and dynamics of DNA-anthracycline complexes. *Biochemistry* **1994,** *33*, 3649-3659.
36. Leonard, G. A.; Hambley, T. W.; McAuley-Hecht, K.; Brown, T.; Hunter, W. N., Anthracycline-DNA interactions at unfavourable base-pair triplet-binding sites: structures of d(CGGCCG)/daunomycin and d(TGGCCA)/adriamycin complexes. *Acta Crystallogr D Biol Crystallogr* **1993,** *49*, 458-467.
37. Gasper, S. M.; Armitage, B.; Shui, X.; Hu, G. G.; Yu, C.; Schuster, G. B.; Williams, L. D., Three-Dimensional Structure and Reactivity of a Photochemical Cleavage Agent Bound to DNA. *Journal of the American Chemical Society* **1998,** *120*, 12402-12409.
38. Nunn, C. M.; Van Meervelt, L.; Zhang, S.; Moore, M. H.; Kennard, O., DNA-drug interactions. *Journal of Molecular Biology* **1991,** *222*, 167-177.

39. Adams, A.; Guss, J. M.; Collyer, C. A.; Denny, W. A.; Prakash, A. S.; Wakelin, L. P. G., Acridinecarboxamide Topoisomerase Poisons: Structural and Kinetic Studies of the DNA Complexes of 5-Substituted 9-Amino-(N-(2-dimethylamino)ethyl)acridine-4-Carboxamides. *Molecular Pharmacology* **2000,** *58*, 649-658.
40. Temperini, C., The crystal structure of the complex between a disaccharide anthracycline and the DNA hexamer d(CGATCG) reveals two different binding sites involving two DNA duplexes. *Nucleic Acids Research* **2003,** *31*, 1464-1469.
41. Gilad, Y.; Senderowitz, H., Docking studies on DNA intercalators. *J. Chem. Inf. Model.* **2014,** *54*, 96-107.
42. Churchill, C. D.; Rutledge, L. R.; Wetmore, S. D., Effects of the biological backbone on stacking interactions at DNA-protein interfaces: the interplay between the backbone...pi and pi...pi components. *Phys. Chem. Chem. Phys.* **2010,** *12*, 14515-14526.
43. Churchill, C. D.; Navarro-Whyte, L.; Rutledge, L. R.; Wetmore, S. D., Effects of the biological backbone on DNA-protein stacking interactions. *Phys. Chem. Chem. Phys.* **2009,** *11*, 10657-10670.
44. Lenz, S. A.; Kellie, J. L.; Wetmore, S. D., Glycosidic bond cleavage in deoxynucleotides: effects of solvent and the DNA phosphate backbone in the computational model. *J. Phys. Chem. B* **2012,** *116*, 14275-14284.
45. Grimme, S., Semiempirical GGA-Type Density Functional Constructed with a Long-Range Dispersion Correction. *J. Comp. Chem.* **2006,** *27*, 1787-1799.
46. Becke, A., Density-Functional Thermochemistry. V. Systematic Optimization of Exchange-Correlation Functionals. *J. Chem. Phys.* **1997,** *107*, 8554-8560.
47. Hehre, W. J.; Stewart, R. F.; Pople, J. A., Self-consistent molecular-orbital methods. I. Use of gaussian expansions of slater-type atomic orbitals. *The Journal of Chemical Physics* **1969,** *51*, 2657-2664.
48. Allen, W. D.; East, A. L. L.; Császár, A. G., In *Structures and Conformations of Non-Rigid Molecules*, Laane, J.; Dakkouri, M.; van der Veken, B.; Oberhammer, H., Eds. Kluwer: Dordrecht, 1993; pp 343-373.
49. Császár, A. G.; Allen, W. D.; Schaefer, H. F., In Pursuit of the ab Initio Limit for Conformational Energy Prototypes *J. Chem. Phys.* **1998,** *108*, 9751-9764.
50. East, A. L. L.; Allen, W. D., The Heat of Formation of Nco. *J. Chem. Phys.* **1993,** *99*, 4638-4650.
51. Weigend, F.; Häser, M.; Patzelt, H.; Ahlrichs, R., RI-MP2: optimized auxiliary basis sets and demonstration of efficiency. *Chemical Physics Letters* **1998,** *294*, 143-152.
52. Pinski, P.; Riplinger, C.; Valeev, E. F.; Neese, F., Sparse maps-A systematic infrastructure for reduced-scaling electronic structure methods. I. An efficient and simple linear scaling local MP2 method that uses an intermediate basis of pair natural orbitals. *J. Chem. Phys.* **2015,** *143*, 034108.
53. Riplinger, C.; Sandhoefer, B.; Hansen, A.; Neese, F., Natural triple excitations in local coupled cluster calculations with pair natural orbitals. *J. Chem. Phys.* **2013,** *139*, 134101.
54. Riplinger, C.; Neese, F., An efficient and near linear scaling pair natural orbital based local coupled cluster method. *J. Chem. Phys.* **2013,** *138*, 034106.
55. Neese, F.; Hansen, A.; Liakos, D. G., Efficient and accurate approximations to the local coupled cluster singles doubles method using a truncated pair natural orbital basis. *J. Chem. Phys.* **2009,** *131*, 064103.

56. Dunning, T. H., Gaussian basis sets for use in correlated molecular calculations. I. The atoms boron through neon and hydrogen. *J. Chem. Phys.* **1989,** *90*, 1007-1023.
57. Feller, D., The Use of Systematic Sequences of Wave Functions for Estimating the Complete Basis Set, Full Configuration Interaction Limit in Water. *J. Chem. Phys.* **1993,** *98*, 7059-7071.
58. Helgaker, T.; Klopper, W.; Koch, H.; Noga, J., Basis-Set Convergence of Correlated Calculations on Water. *J. Chem. Phys.* **1997,** *106*, 9639-9646.
59. Kruse, H.; Banáš, P.; Šponer, J., Investigations of Stacked DNA Base-Pair Steps: Highly Accurate Stacking Interaction Energies, Energy Decomposition, and Many-Body Stacking Effects. *Journal of Chemical Theory and Computation* **2019,** *15*, 95-115.
60. Becke, A. D., Density-functional thermochemistry. III. The role of exact exchange. *J. Chem. Phys.* **1993,** *98*, 5648-5652.
61. Grimme, S.; Antony, J.; Ehrlich, S.; Krieg, H., A consistent and accurate ab initio parametrization of density functional dispersion correction (DFT-D) for the 94 elements H-Pu. *Journal of Chemical Physics* **2010,** *132*, 154104.
62. Zhao, Y.; Truhlar, D. G., The M06 suite of density functionals for main group thermochemistry, thermochemical kinetics, noncovalent interactions, excited states, and transition elements: two new functionals and systematic testing of four M06-class functionals and 12 other functionals. *Theor. Chem. Acc.* **2008,** *120*, 215-241.
63. Chai, J. D.; Head-Gordon, M., Long-range corrected hybrid density functionals with damped atom-atom dispersion corrections. *Phys. Chem. Chem. Phys.* **2008,** *10*, 6615-6620.
64. Weigend, F.; Ahlrichs, R., Balanced basis sets of split valence, triple zeta valence and quadruple zeta valence quality for H to Rn: Design and assessment of accuracy. *Phys. Chem. Chem. Phys.* **2005,** *7*, 3297-3305.
65. Grimme, S.; Ehrlich, S.; Goerigk, L., Effect of the damping function in dispersion corrected density functional theory. *J. Comput. Chem.* **2011,** *32*, 1456-1465.
66. Sure, R.; Grimme, S., Corrected small basis set Hartree-Fock method for large systems. *Journal of Computational Chemistry* **2013,** *34*, 1672-1685.
67. Grimme, S.; Brandenburg, J. G.; Bannwarth, C.; Hansen, A., Consistent structures and interactions by density functional theory with small atomic orbital basis sets. *Journal of Chemical Physics* **2015,** *143*, 054107.
68. E. D. Glendening, A. E. R., J. E. Carpenter, F. Weinhold, NBO Version 3.1.
69. Reed, A. E.; Weinstock, R. B.; Weinhold, F., Natural-Population Analysis. *J. Chem. Phys.* **1985,** *83*, 735-746.
70. Frisch, M.; Trucks, G.; Schlegel, H.; Scuseria, G.; Robb, M.; Cheeseman, J.; Scalmani, G.; Barone, V.; Mennucci, B.; Petersson, G.; Nakatsuji, H.; Caricato, M.; Li, X.; Hratchian, H.; Izmaylov, A.; Bloino, J.; Zheng, G.; Sonnenberg, J.; Hada, M.; Ehara, M.; Toyota, K.; Fukuda, R.; Hasegawa, J.; Ishida, M.; Nakajima, T.; Honda, Y.; Kitao, O.; Nakai, H.; Vreven, T.; Montgomery, J., JA; Peralta, J.; Ogliaro, F.; Bearpark, M.; Heyd, J.; Brothers, E.; Kudin, K.; Staroverov, V.; Keith, T.; Kobayashi, R.; Normand, J.; Raghavachari, K.; Rendell, A.; Burant, J.; Iyengar, S.; Tomasi, J.; Cossi, M.; Rega, N.; Millam, J.; Klene, M.; Knox, J.; Cross, J.; Bakken, V.; Adamo, C.; ramillo, J.; Gomperts, R.; Stratmann, R.; Yazyev, O.; Austin, A.; Cammi, R.; Pomelli, C.; Ochterski, J.; Martin, R.; Morokuma, K.; Zakrzewski, V.; Voth, G.; Salvador, P.; Dannenberg, J.; Dapprich, S.; Daniels, A.; Farkas, O.; Foresman, J.; Ortiz, J.; Cioslowski, J.; Fox, D. *Gaussian 09, Revision D.01*, Gaussian, Inc.: 2009.

71. Neese, F., Software update: the ORCA program system, version 4.0. *Wiley Interdisciplinary Reviews-Computational Molecular Science* **2018,** *8*.