The Journal of Chemical Physics

NENCI-2021. I. A large benchmark database of non-equilibrium non-covalent interactions emphasizing close intermolecular contacts

Cite as: J. Chem. Phys. **155**, 184303 (2021); https://doi.org/10.1063/5.0068862
Submitted: 27 August 2021 • Accepted: 20 September 2021 • Published Online: 11 November 2021

Zachary M. Sparrow, Brian G. Ernst, Paul T. Joo, et al.

COLLECTIONS

Paper published as part of the special topic on JCP Editors' Choice 2021

F This paper was selected as Featured

View Online   Export Citation   CrossMark

ARTICLES YOU MAY BE INTERESTED IN

Software for the frontiers of quantum chemistry: An overview of developments in the Q-Chem 5 package
The Journal of Chemical Physics **155**, 084801 (2021); https://doi.org/10.1063/5.0055522

A consistent and accurate ab initio parametrization of density functional dispersion correction (DFT-D) for the 94 elements H-Pu
The Journal of Chemical Physics **132**, 154104 (2010); https://doi.org/10.1063/1.3382344

r^2SCAN-3c: A "Swiss army knife" composite electronic-structure method
The Journal of Chemical Physics **154**, 064103 (2021); https://doi.org/10.1063/5.0040021

AIP Publishing

NENCI-2021. I. A large benchmark database of non-equilibrium non-covalent interactions emphasizing close intermolecular contacts

Cite as: J. Chem. Phys. **155**, 184303 (2021); doi: 10.1063/5.0068862
Submitted: 27 August 2021 • Accepted: 20 September 2021 •
Published Online: 11 November 2021

Zachary M. Sparrow, Brian G. Ernst, Paul T. Joo, Ka Un Lao, and Robert A. DiStasio, Jr.[a)]

AFFILIATIONS

Department of Chemistry and Chemical Biology, Cornell University, Ithaca, New York 14853, USA

[a)] Author to whom correspondence should be addressed: distasio@cornell.edu

ABSTRACT

In this work, we present NENCI-2021, a benchmark database of ~8000 Non-Equilibirum Non-Covalent Interaction energies for a large and diverse selection of intermolecular complexes of biological and chemical relevance. To meet the growing demand for large and high-quality quantum mechanical data in the chemical sciences, NENCI-2021 starts with the 101 molecular dimers in the widely used S66 and S101 databases and extends the scope of these works by (i) including 40 cation–π and anion–π complexes, a fundamentally important class of non-covalent interactions that are found throughout nature and pose a substantial challenge to theory, and (ii) systematically sampling all 141 intermolecular potential energy surfaces (PESs) by *simultaneously* varying the intermolecular distance and intermolecular angle in each dimer. Designed with an emphasis on close contacts, the complexes in NENCI-2021 were generated by sampling seven intermolecular distances along each PES (ranging from 0.7× to 1.1× the equilibrium separation) and nine intermolecular angles per distance (five for each ion–π complex), yielding an extensive database of 7763 benchmark intermolecular interaction energies (E_{int}) obtained at the coupled-cluster with singles, doubles, and perturbative triples/complete basis set [CCSD(T)/CBS] level of theory. The E_{int} values in NENCI-2021 span a total of 225.3 kcal/mol, ranging from −38.5 to +186.8 kcal/mol, with a mean (median) E_{int} value of −1.06 kcal/mol (−2.39 kcal/mol). In addition, a wide range of intermolecular atom-pair distances are also present in NENCI-2021, where close intermolecular contacts involving atoms that are located within the so-called van der Waals envelope are prevalent—these interactions, in particular, pose an enormous challenge for molecular modeling and are observed in many important chemical and biological systems. A detailed symmetry-adapted perturbation theory (SAPT)-based energy decomposition analysis also confirms the diverse and comprehensive nature of the intermolecular binding motifs present in NENCI-2021, which now includes a significant number of primarily induction-bound dimers (e.g., cation–π complexes). NENCI-2021 thus spans all regions of the SAPT ternary diagram, thereby warranting a new four-category classification scheme that includes complexes primarily bound by electrostatics (3499), induction (700), dispersion (1372), or mixtures thereof (2192). A critical error analysis performed on a representative set of intermolecular complexes in NENCI-2021 demonstrates that the E_{int} values provided herein have an average error of ±0.1 kcal/mol, even for complexes with strongly repulsive E_{int} values, and maximum errors of ±0.2–0.3 kcal/mol (i.e., ~±1.0 kJ/mol) for the most challenging cases. For these reasons, we expect that NENCI-2021 will play an important role in the testing, training, and development of next-generation classical and polarizable force fields, density functional theory approximations, wavefunction theory methods, and machine learning based intra- and inter-molecular potentials.

I. INTRODUCTION

With tunable strengths situated between thermal fluctuations and covalent bonds, non-covalent interactions (NCIs) are ubiquitous in nature and play a critical role in determining the structure, stability, and function in a number of systems throughout chemistry, biology, physics, and materials science.[1–5] One particularly illustrative example is the famous DNA double helix, whose structure is stabilized by a complex network of hydrogen bonds and π–π stacking interactions between constituent nucleobases. In organic synthesis and biochemistry, many catalysts and enzymes function by leveraging NCIs to position/orient substrates for the ensuing reaction and/or stabilize critical points along the reaction pathway, e.g., ion–π interactions can stabilize intermediates and transition states with excess charge.[6–11] Over the past two decades, NCIs have garnered critical recognition throughout the chemical sciences and have now become an integral part of "chemical intuition" when rationalizing complex chemical structures and/or processes and designing molecular systems (e.g., catalysts) for optimal performance and/or novel applications. In this regard, there are quite a number of NCI-based applications actively under investigation, ranging from crystal engineering (where hydrogen and halogen bonding are used to direct molecular assembly)[12,13] and artificial molecular machines [where π–π stacking, hydrogen bonding, and dispersion/van der Waals (vdW) forces are leveraged to control complex motion at the nanoscale][14–17] to drug discovery (where candidate molecules are selected and screened based on specific NCIs present in the corresponding active site).[18–21]

Given their importance and prevalence, it is imperative that there exists a suite of computational methods that can provide an accurate, reliable, and computationally efficient description of NCIs for systems ranging from small gas-phase molecular dimers to the complex tertiary and quaternary structure of proteins in solvent.[22] To meet these goals, a number of computational techniques have been developed over the past century, including (but not limited to) model intermolecular potentials (e.g., Lennard-Jones),[23] classical and polarizable force fields,[24,25] density functional theory (DFT) approximations with corrections for dispersion/vdW interactions,[26–28] efficient (i.e., linear scaling) algorithms for highly accurate wavefunction theory (WFT) methods,[29,30] and, more recently, the large and rapidly growing suite of machine learning (ML) based approaches.[31–35] During this time, such computational methods have enjoyed tremendous success and made critical contributions to a number of different fields, e.g., identifying promising pharmaceutical molecules,[18–21] predicting (meta-)stable molecular crystal polymorphs,[36,37] and elucidating supercritical behavior in high-pressure liquid hydrogen,[38] to name a few. However, we would argue that the next generation of theoretical approaches for describing NCIs would tremendously benefit by addressing the following challenges in an accurate, reliable, and computationally efficient manner: (i) the need to describe NCIs in large molecular and condensed-phase systems (i.e., collective many-body effects, solvation/solvent effects, and simultaneous treatment of short-, intermediate-, and long-range NCIs on the same footing);[39–44] (ii) the need to describe the diverse types of NCIs on the same footing (i.e., similar performance for hydrogen bonding, π–π stacking, dispersion, ion–π interactions, etc.);[45–48] and (iii) the need to describe NCIs in equilibrium and non-equilibrium systems on the same footing [i.e., similar performance across entire potential energy surfaces (PESs)].[49,50]

An essential part of developing next-generation theoretical methods for describing NCIs involves testing and/or training new approximations against highly accurate benchmark data. For the increasingly popular suite of ML-based models—which require large amounts of high-quality data to learn the quantum mechanics underlying NCIs—such reference data are of critical importance. However, such benchmark non-covalent/non-bonded interaction (or binding) energies (E_{int}) are seldom experimentally available, especially for large/complex systems and non-equilibrium configurations. Instead, one usually relies on quantum chemical and/or quantum Monte Carlo methods (i.e., WFT methods) to obtain highly accurate and systematically improvable E_{int} values for benchmarking and training purposes.[51] On the WFT side, coupled-cluster theory, including single, double, and perturbative triple excitations, in conjunction with an extrapolation to the complete basis set limit [CCSD(T)/CBS] has long been considered the *de facto* "gold standard" for generating accurate E_{int} data for small- and medium-sized organic molecules, and has therefore been used to generate a number of seminal benchmark databases for NCIs.[52,53] One of the first of these databases, the so-called S22 database,[54,55] includes 22 CCSD(T)-quality E_{int} values for a set of small-/medium-sized biologically relevant intermolecular complexes (comprised of {C, H, O, N}) in their respective optimized (equilibrium) geometries, and was designed to cover a number of different intermolecular binding motifs (i.e., single and double hydrogen bonds, dipole–dipole interactions, π–π stacking, dispersion, C–H$\cdots\pi$, etc.). Following the success and widespread use of S22 in the testing and parameterization of many theoretical methods for describing NCIs, the amount of benchmark-quality E_{int} data was substantially increased with the introduction of the S66 database (which includes 66 equilibrium intermolecular complexes of similar size and composition to that found in S22) and extensions thereof to include complexes with non-equilibrium intermolecular distances (along a series of dissociation curves in S22x5[56] and S66x8[57,58]) and non-equilibrium intermolecular angles (at the equilibrium distance in S66a8).[59] During the same time, other benchmark NCI databases were constructed to reflect the diverse number of NCI types (or binding motifs) found in halogen-containing systems (X40x10),[60] nucleobase dimers (ACHC),[61] charge transfer (CT) complexes,[62] alkane dimers (ADIM6),[63] large molecular dimers (L7),[42] host–guest complexes (S12L),[39] halogen-bonded systems (XB18),[64] sulfur-containing systems (SULFURx8),[65] and many more.[52,55,66–71] Along similar lines, there are also NCI databases based on a symmetry-adapted perturbation theory (SAPT) decomposition of E_{int} into components (i.e., electrostatics, exchange, induction, and dispersion), which have been used to train force fields for molecular dynamics simulations.[72–74] Of particular interest here is the S101x7 database,[49] which starts with the molecular dimers in S66 and expands this set to include 35 additional biologically relevant complexes containing halogens (i.e., F, Cl, and Br) and second-row elements (i.e., S and P), as well as additional intermolecular complexes involving charged systems and/or water. Like the S66x8 database, S101x7 also includes complexes with non-equilibrium intermolecular distances by computing SAPT-based E_{int} values for select points along each intermolecular PES; in the S101x7 case, these seven points ranged from 0.7× to 1.1×, the equilibrium intermolecular separation in an effort to better capture short-range charge penetration effects.[49] For additional information about benchmark databases and the methodologies used to construct them, we also point the reader to the review by Řezáč and Hobza.[75]

While existing databases are growing in size, most are still relatively small (containing $\lesssim 500$ interaction energies), making them insufficient for the rapidly growing field of ML-based intra-/inter-molecular potentials (in particular, for training truly deep learning models). In this regard, composite databases (such as GMTKN55,[70] ACCDB,[76] and NCIAtlas[77–79]) and the very recently published datasets of Donchev *et al.*[80] (DES15K/DES370K) can be considerably larger and represent a key step toward meeting such growing data demands. However, the use of different basis sets and/or quantum chemical methods during the generation and compilation of some of these data may potentially be a source of both random and systematic errors in some ML applications. Due to the high computational cost of generating benchmark E_{int} values for large systems, most existing databases (with the exception of L7[42] and S12L[39]) have been limited to small-to-medium organic/biological molecules (usually containing < 20 atoms); as a consequence, many of these databases do not capture the collective nature of NCIs (i.e., many-body effects, solvation/solvent effects, and NCIs across multiple length scales) present in large/complex molecules and condensed-phase systems. In addition, most existing databases have focused on common intermolecular binding motifs, such as hydrogen and halogen bonding, π–π stacking, dipole–dipole interactions, dispersion, and C–H$\cdots\pi$ interactions, while other important binding motifs (such as cation–π and anion–π interactions) have been largely underrepresented. As such, these databases tend to include intermolecular complexes that are primarily bound by electrostatics, dispersion, or a mixture thereof, but have not included intermolecular complexes that are primarily induction-bound. Furthermore, prior databases (e.g., S22x5, S66x8, S66a8, and S101x7) primarily focused on the equilibrium geometry and a *single* displacement from the equilibrium geometry (i.e., scaling the intermolecular distance or rotating one monomer), but very few have explored wider swaths of the intermolecular PES. In this regard, most databases have also only slightly touched upon close intermolecular contacts (i.e., the short-range and often repulsive sector of the intermolecular PES), although there are some examples where such short-range considerations have been incorporated (e.g., S101x7,[49] R160x6,[81] and R739x5[79]). As a result, the performance of many theoretical methods for accurately and reliably describing NCIs in large and complex systems, for a diverse array of binding motifs, and across significant portions of the intermolecular PES is simply not well known.

Accurate and reliable descriptions of non-equilibrium NCIs at reduced intermolecular separations—where several strong and competing short-range intermolecular forces are at play—are important for a number of reasons and pose a substantial challenge to theory. For instance, there are numerous examples throughout chemistry and chemical biology where close intermolecular contacts are either present at equilibrium or force the system to adopt a different configuration.[82,83] A striking example of this was recently observed when studying the enantioselectivity of sBOX catalysts, where a combination of attractive and repulsive NCIs is responsible for the enantio-determining C–CN bond formation in chiral nitriles.[83] Intermolecular close contacts also play a crucial role in the study of systems operating under high-pressure conditions, ranging from the microscopic structure of supercritical water[84] to the high-pressure synthesis of compounds with atypical compositions and novel properties,[85] and the search for high-T_c superconducting materials.[86] Theoretically speaking, SAPT decomposition studies[87] have shown that the intermolecular distance can have a profound influence on the absolute and relative magnitudes of the underlying E_{int} components (i.e., electrostatics, exchange, induction, and dispersion), implying that the forces present in short-range non-equilibrium intermolecular complexes of small/simpler molecules can mimic those found in larger/more complicated systems at equilibrium separations. Interestingly, this also suggests that training and/or testing theoretical methods on non-equilibrium configurations (particularly in the short-range) of small-to-medium molecular dimers can be used as a surrogate for describing the NCIs in a more diverse range of large (and possibly intractable) systems. Since finite-temperature molecular dynamics and Monte Carlo simulations require a consistent treatment of the structures and energetics across the entire PES, an accurate and reliable treatment of non-equilibrium NCIs (including short-range as well as intermediate- and long-range interactions) is of enormous importance for these applications as well. However, the difficulties in obtaining such an accurate and reliable theoretical description of non-equilibrium NCIs across multiple length scales should also be emphasized. For instance, the long-range sector of the intermolecular PES requires a balanced description of both electrostatics and dispersion, and this can be particularly challenging when dealing with NCIs that also include charged species and/or molecules with substantial multipole moments. For larger intermolecular separations, intermolecular energies (and forces) tend to be small, which provides additional challenges when trying to describe points along the PES on the same footing. At reduced intermolecular separations, the increased amount of orbital (or density) overlap between monomers gives rise to a complex interplay between strongly attractive and strongly repulsive intermolecular forces (e.g., charge transfer and penetration, Pauli repulsion, and many-body exchange–correlation effects), and an error when describing any one of these components can lead to disastrous results.[49,88,89] For such short-range non-equilibrium NCIs, the performance of the current suite of theoretical methods is still an open question, and a number of studies have reported higher errors for repulsive intermolecular contacts.[55,81,89,90] In this regime, even the suitability of high-level WFT-based approaches for generating benchmark E_{int} data is still largely unresolved as such approaches suffer from issues related to the use of incomplete basis sets (i.e., basis set incompleteness and superposition errors) in conjunction with an approximate treatment of electron correlation effects (including questions regarding the reliability of perturbative expansions).

In this work, we directly address the aforementioned challenges needed for training, testing, and developing next-generation theoretical approaches for describing NCIs by introducing NENCI-2021, a benchmark database of ~8000 Non-Equilibirum Non-Covalent Interaction energies for a diverse selection of 141 molecular dimers of biological and chemical relevance. Starting with the 101 dimers in the S101[49] (and hence S66[57]) databases, which contain a diverse set of intermolecular binding motifs (i.e., single and double hydrogen bonds, halogen bonds, ion–dipole and dipole–dipole interactions, π–π stacking, dispersion, and X–H$\cdots\pi$) and a large number of molecular dimers involving water (which represents a crucial first step toward generating benchmark E_{int} values in aqueous environments), NENCI-2021 extends the scope of these seminal works in two directions. For one, NENCI-2021 includes 40 cation–π and anion–π complexes, a fundamentally important and particularly strong class of NCIs that are primarily induction-bound[48] and characterized by equilibrium E_{int} values which are typically larger in magnitude than hydrogen bonds and salt bridges. As such, an accurate and reliable description of ion–π interactions poses substantial difficulties for theory, and their inclusion in NENCI-2021 directly addresses the challenge of simultaneously describing diverse NCI types on the same footing [i.e., point (*ii*) above]. Second, NENCI-2021 also includes an extensive and systematic sampling of equilibrium and non-equilibrium configurations on each of the 141 intermolecular PESs by *simultaneously* varying the intermolecular distance and intermolecular angle in each dimer. Designed with an emphasis on close intermolecular contacts, the complexes in NENCI-2021 were generated by sampling seven intermolecular distances (ranging from 0.7× to 1.1× the equilibrium separation) and nine intermolecular angles per distance (five for each ion–π complex), yielding an extensive database of 7763 benchmark E_{int} values obtained at the CCSD(T)/CBS level of theory. In doing so, NENCI-2021 directly addresses the challenges of describing the collective nature of NCIs in large/complex systems [i.e., point (i)] and simultaneously describing NCIs in equilibrium and non-equilibrium systems on the same footing [i.e., point (iii)]. Of these 7763 intermolecular complexes, 6363 are derived from the molecular dimers included in the S101 database[49] (which includes a total of 2079 complexes involving water, 63 of which are water dimers) and 1400 are newly added ion–π complexes. All the intermolecular complexes in NENCI-2021 contain at least one first-row element (C, N, and O), 1715 intermolecular complexes contain a halogen (F, Cl, and Br), 693 contain at least one second-row main-group element (S and P, not including Cl), and 700 contain an alkali-metal ion (Li^+ and Na^+). The E_{int} values in NENCI-2021 span a total of 225.3 kcal/mol, ranging from −38.5 kcal/mol (corresponding to the strongly attractive $Li^+ \cdots$ benzene ion–π complex) to +186.8 kcal/mol (corresponding to a strongly repulsive DMSO⋯DMSO complex that has been scaled to 0.7× the equilibrium intermolecular separation and rotated to a non-equilibrium angle), with a mean (median) E_{int} value of −1.06 kcal/mol (−2.39 kcal/mol). A detailed SAPT-based energy decomposition analysis demonstrates the diverse and comprehensive nature of NENCI-2021, which spans all regions of the corresponding ternary diagram and includes intermolecular binding motifs primarily bound by electrostatics (3499), induction (700), dispersion (1372), or mixtures thereof (2192). A critical error analysis performed on a representative set of intermolecular complexes in NENCI-2021 demonstrates that the E_{int} values provided herein at the CCSD(T)/CBS level have an average error of ±0.1 kcal/mol, even for complexes with strongly repulsive E_{int} values, and maximum errors of ±0.2–0.3 kcal/mol (i.e., ∼±1.0 kJ/mol) for the most challenging cases.

When compared to the current state of the art in benchmark NCI databases, namely, NCIAtlas[77–79] and DES15K[80] (a subset of the larger DES370K database with a PES resolution more comparable to NENCI-2021 and NCIAtlas), NENCI-2021 is roughly the same size and computed at a comparable level of theory [i.e., CCSD(T) extrapolated to the CBS limit]. However, NCIAtlas and DES15K primarily focus on one-dimensional (either radial or angular) scans of the intermolecular PES, while NENCI-2021 contains two-dimensional (simultaneous radial and angular) scans of the PES that more thoroughly sample close intermolecular contacts. Although ion–π interactions are among the strongest NCIs known, NCIAtlas excludes this class of interactions completely and DES15K only contains cation–π interactions, while NENCI-2021 includes a quite substantial number (1400) of both cation–π and anion–π complexes. Designed to meet the growing demand for large and high-quality quantum mechanical data in the chemical sciences, we expect that NENCI-2021 will complement these state-of-the-art benchmark NCI databases and be another important resource for testing, training, and developing next-generation force fields, DFT approximations, WFT methods, and ML-based intra-/inter-molecular potentials.

The remainder of this article is organized as follows. Section II describes the construction of NENCI-2021, including the selection of molecular dimers, generation of equilibrium and non-equilibrium intermolecular complexes, a detailed description of the employed computational protocol, and a guide to obtaining the database. Section III discusses the properties of NENCI-2021, including a statistical analysis of the intermolecular interaction energies and closest intermolecular contacts, an SAPT-based energy decomposition analysis of the intermolecular binding motifs, and a critical assessment of the error in the benchmark E_{int} values provided herein. This article ends with some brief conclusions and future directions in Sec. IV. In a follow-up to this work,[91] many popular WFT and DFT methods are explicitly tested on the NENCI-2021 database, where it is shown that there is a nearly universal increase in error when describing the repulsive wall of the intermolecular PES and that ion–π complexes can be quite challenging to model in an accurate and reliable fashion.

II. CONSTRUCTION OF THE NENCI-2021 DATABASE

A. Selection of molecular dimers

NENCI-2021 is a large database of ∼8000 benchmark intermolecular interaction energies (E_{int}; see Sec. II C) that includes a diverse selection of molecular dimers and binding motifs of biological and chemical relevance, with an emphasis on non-equilibrium (attractive and repulsive) configurations and close intermolecular contacts. As depicted in the left panel of Fig. 1, the construction of NENCI-2021 starts with the 101 molecular dimers in the S101[49] database (a superset containing the earlier constructed S66[57] database), which were carefully chosen to contain small molecules with the NCIs found in biological and chemical systems. As such, NENCI-2021 inherits the extensive sampling of molecule types in S66 and S101, which are comprised of the {H, C, N, O, F, P, S, Cl, Br} atom types, ranging in size from small (e.g., H_2O, ethene, and ethyne) to medium (e.g., uracil, indole, and pentane), and include second- and third-row elements [e.g., dimethyl sulfoxide (DMSO), MeCl, and BenBr] and positively (e.g., $MeNH_3^+$, imidazole$^+$, and guanidine$^+$) and negatively (e.g., AcO^-, $H_2PO_4^-$, and HPO_4^{2-}) charged species. In addition, NENCI-2021 also inherits a wide variety of intermolecular binding motifs, including dimers with single and double hydrogen bonds, halogen bonds,[92] and X–H$\cdots\pi$ interactions, as well as intermolecular complexes primarily bound by dispersion, electrostatics (e.g., ion–dipole and dipole–dipole), and mixtures thereof. Another salient benefit of using S66 and S101 as the foundation for NENCI-2021 is the large number of dimers involving water, which provides a crucial first step toward the generation of benchmark intermolecular interaction energies in an aqueous environment.

NENCI-2021 extends these databases in the following two ways: (i) it includes 40 new cation–π and anion–π complexes for a total of 141 molecular dimers and (ii) it systematically samples both equilibrium and non-equilibrium intermolecular distances and intermolecular angles for each dimer (with a particular emphasis on close intermolecular contacts) for a total of 7763 benchmark interaction energies. In particular, NENCI-2021 includes ion–π

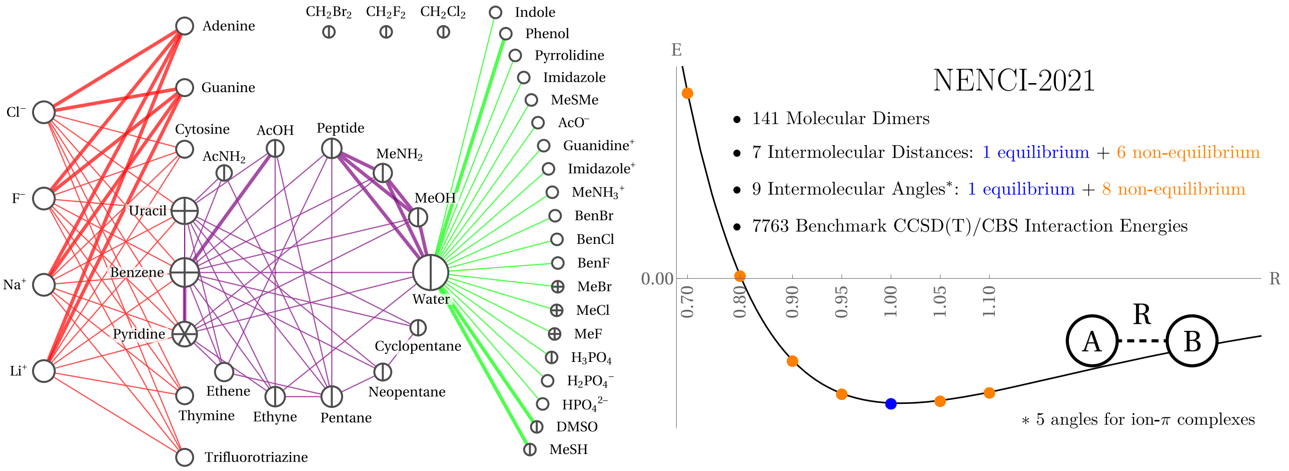

FIG. 1. (*left*) Graphical depiction of the 141 molecular dimers in the NENCI-2021 database. NENCI-2021 contains all molecular dimers in the original S66 database[57] (purple lines), the additional dimers present in the S101 (superset) database[49] (green lines, S66 ⊂ S101), and a new set of 40 cation–π and anion–π complexes (red lines, S66 ⊂ S101 ⊂ NENCI-2021). In this graph, each monomer is represented by a vertex, the size of which is proportional to the number of molecular dimers involving that monomer; graph edges connecting two vertices indicate a molecular dimer formed from the connected monomers. The bold edges between vertices denote two different molecular dimer orientations involving the connected monomers (e.g., for water–phenol, water is the hydrogen-bond donor in one dimer and the hydrogen-bond acceptor in the other). Chords passing through the center of a vertex indicate a molecular dimer formed from a single monomer (e.g., there is one water dimer, two uracil dimers, and three pyridine dimers in NENCI-2021). (*right*) Overall description of the NENCI-2021 database. For each of the 141 dimers described above, NENCI-2021 generates a series of equilibrium and non-equilibrium configurations by *simultaneously* sampling seven intermolecular distances and nine intermolecular angles (five for the ion–π complexes due to symmetry considerations). As such, NENCI-2021 includes 7763 benchmark CCSD(T)/CBS intermolecular interaction energies, which correspond to a wide range of equilibrium and non-equilibrium (both repulsive and attractive) geometries and emphasize close intermolecular contacts. See Secs. II A–II C for the details regarding the construction of NENCI-2021.

complexes comprised of the simplest biologically relevant monovalent cations (Li^+ and Na^+) and anions (F^- and Cl^-) interacting with a representative set of π-systems, which includes the five DNA/RNA nucleobases (adenine, cytosine, guanine, thymine, and uracil), benzene, pyridine, and trifluorotriazine. The inclusion of ion–π complexes in NENCI-2021 was primarily driven by the fact that ion–π interactions are among the strongest NCIs known (with intermolecular interaction energies often rivaling that of hydrogen bonds and salt bridges) and have been observed throughout chemistry and biology.[93–97] This extension was also motivated by some of our recent work,[48] which used SAPT[88] to demonstrate that cation–π complexes are primarily bound by induction, while anion–π complexes are bound by a complex interplay between induction, dispersion, and electrostatics; as such, their inclusion substantially expands the scope/range of intermolecular binding motifs in NENCI-2021 (see Sec. III B). As shown in Paper II[91] of this series, this complex interplay between intermolecular forces (in addition to the presence of charged atomic species) in ion–π complexes poses a unique challenge when trying to obtain accurate and reliable intermolecular interaction energies using both WFT and DFT methods. In addition, the inclusion of promiscuous ion–π binders (i.e., π-systems such as the DNA/RNA nucleobases, which can form favorable ion–π complexes with both cations and anions[48]) and π-systems that can only form energetically favorable ion–π complexes with cations (e.g., benzene) or anions (e.g., trifluorotriazine) is also well-aligned with one of the fundamental goals of NENCI-2021, i.e., to provide a more comprehensive sampling of both attractive and repulsive non-equilibrium configurations containing a diverse array of NCI types.

Motivated by the S22x5,[56] S66x8,[57,58] and S101x7[49] databases, in which intermolecular interaction energy curves were constructed for each molecular dimer, and the S66a8[59] database, in which intermolecular angles were sampled, NENCI-2021 systematically samples both equilibrium and non-equilibrium intermolecular distances and intermolecular angles for each of the 141 molecular dimers described above. As depicted in the right panel of Fig. 1, NENCI-2021 samples seven intermolecular distances (i.e., 0.7×, 0.8×, 0.9×, 0.95×, 1.0×, 1.05×, 1.1× the equilibrium intermolecular separation) and nine intermolecular angles (only five intermolecular angles for the ion–π complexes, *vide infra*); for more details, see Sec. II B. NENCI-2021 therefore contains benchmark intermolecular interaction energies (see Secs. II C and III C) for $7 \times 9 = 63$ geometries (configurations) for each of the 101 molecular dimers in the S101 database and $7 \times 5 = 35$ geometries for each of the 40 ion–π complexes, yielding a total of $63 \times 101 + 35 \times 40 = 7763$ equilibrium and non-equilibrium intermolecular complexes. By including such a systematic sampling of equilibrium and non-equilibrium structures, NENCI-2021 is a relatively large database that contains a wide range of attractive and repulsive intermolecular interaction energies (see Sec. III A); as such, we believe that NENCI-2021 will be well-suited for in-depth studies of the NCIs found throughout biology and chemistry, as well as training and testing next-generation density functional approximations, dispersion corrections, polarizable force fields, and ML-based potentials. By including an extensive set of angularly sampled geometries at 0.7× and 0.8× the equilibrium intermolecular separation, NENCI-2021 also includes a wide range of close intermolecular contacts, which are found throughout chemistry and chemical biology as well as high-pressure systems; here, we stress again that benchmark intermolecular interaction energies in this regime not only serve as surrogates for larger/more complex systems at equilibrium, but are also important to ensure similar performance across the entire intermolecular PES when training, testing, and developing novel theoretical methods.

B. Generation of equilibrium and non-equilibrium intermolecular complexes

Unless otherwise specified, all monomer geometries were taken from the S66[57] and S101[49] databases. For the eight π-systems used to construct the 40 ion–π complexes in NENCI-2021, the monomer geometries for benzene, pyridine, and uracil were taken from the S66[57] database, while the monomer geometries for trifluorotriazine and the DNA nucleobases were taken from our recent work on promiscuous ion–π binding.[48] During the construction of the equilibrium and non-equilibrium molecular dimer geometries, we employed the frozen monomer convention in which all monomers were kept fixed at their optimized geometries.

The 66 molecular dimer geometries in the S66[57] database were also taken as is and without any changes; for the remaining 75 molecular dimers, equilibrium geometries were optimized (see Sec. II C) along a pre-defined characteristic intermolecular interaction vector. This characteristic intermolecular interaction vector was based on the interaction type (e.g., hydrogen-bonded, halogen-bonded, dispersion-bound, and ion–π) assigned to the molecular dimer via chemical intuition. Dimers that appear in the S66 database were assigned the same interaction type as in the original work,[57,59] and the remaining dimers were assigned an interaction type that was as consistent as possible with the S66 convention. Given one of the following interaction types, the characteristic intermolecular interaction vector was defined as follows:

- For hydrogen (halogen)-bonded systems, the interaction vector points between the hydrogen (halogen) bond donor and the hydrogen (halogen) bond acceptor. For double-hydrogen-bonded systems, the interaction vector is defined as the mean of the two hydrogen-bond vectors (with both taken to originate from the same monomer).
- For dispersion-bound systems, the interaction vector points from the center of mass of monomer *A* to the center of mass of monomer *B*.
- For ion–π complexes, the interaction vector points from the ion to the nuclear center of charge of the π-system (computed using only the atoms in each ring, i.e., the five carbons and nitrogen in pyridine). Here, we note in passing that this on-axis placement of the ion does not necessarily correspond to the lowest-energy geometry of each ion–π complex.[98,99]
- Finally, there remain a few special cases (i.e., the T-shaped benzene dimer), which do not fit well into any of these categories. Such systems are treated analogously with the dispersion-bound complexes, but only a subset of atoms is used in calculating an effective "molecular center" to ensure that the interaction vector accurately characterizes the interaction. For reference, the atoms used to calculate the interaction vector for each such complex are provided in Table S1.

All remaining equilibrium molecular dimer geometries were obtained by minimizing the intermolecular interaction energy by rigidly translating monomer *A* along the characteristic intermolecular interaction vector (see Sec. II C) and then used as starting points to generate all non-equilibrium structures.

To systematically sample both intermolecular distances and intermolecular angles for the 141 molecular dimers in NENCI-2021, we started with the procedure devised by Řezáč, Riley, and Hobza when constructing the S66x8[57] and S66a8[59] databases, and extended this protocol to accommodate a broader range of intermolecular interaction types and orientations. As such, the 528 molecular dimer geometries in the S66a8[59] database were also taken as is and without any changes. The procedure for generating the remaining 7094 non-equilibrium intermolecular complexes in NENCI-2021 is outlined below, with steps 1–5 graphically illustrated for the water dimer in Fig. 2.

STEP 1. Starting with an optimized equilibrium intermolecular complex, arbitrarily label each monomer as either *A* or *B* (except for the ion–π complexes, in which the ion should be labeled monomer *A*). Draw the characteristic intermolecular interaction vector from *B* to *A* (black dashed line) according to the interaction type assigned to the molecular dimer (*vide supra*). Define the *z* axis (solid red arrow) along the interaction vector.

STEP 2. Without loss of generality, assume that *A* will be rotated around *B* (the alternative will be dealt with in step 8 below). To determine the axes of rotation, first find the principal axis (C_n) corresponding to monomer *A* [i.e., the molecular axis with the highest degree (n) of rotational symmetry]; for the water monomer depicted in Fig. 2, the principal axis is the black solid line labeled C_2. If no principal axis with $n \geq 2$ exists, we follow the convention used during the construction of the S66a8[59] database, i.e., an approximate principal axis is defined by removing all hydrogen atoms from the molecule and reducing the identity of each heavy atom and functional group to identical spheres.

STEP 3. Define the *y* axis (yellow solid arrow) to be perpendicular to the *z* axis and the principal axis of *A*.

STEP 4. Define the *x* axis (blue solid arrow) to be perpendicular to the *z* and *y* axes, thereby completely specifying the local reference frame used in this work.

STEP 5. To generate preliminary geometries for the first four non-equilibrium intermolecular angles, rotate *A* about the *x* and *y* axes passing through the tail of the interaction vector (i.e., located on monomer *B*) by $\theta = \pm 30^\circ$.

STEP 6. For each non-equilibrium intermolecular angle, minimize the intermolecular interaction energy by rigidly translating *A* along the characteristic intermolecular interaction vector (see Sec. II C). For the ion–π complexes that are repulsive along the entire dissociation curve (e.g., $Na^+ \cdots$ trifluorotriazine), the minimum of the SAPT exchange + induction + dispersion (EID) energy[48] was used *in lieu* of the intermolecular interaction energy (see Sec. II C). Define the intermolecular distance (i.e., the length of the characteristic intermolecular interaction vector) in each optimized geometry as the equilibrium (1.0×) intermolecular distance for the given non-equilibrium intermolecular angle.

STEP 7. For each non-equilibrium intermolecular angle, scale the corresponding (optimized) interaction vector by factors of 0.7×, 0.8×, 0.9×, 0.95×, 1.05×, and 1.1×, and rigidly translate *A* consistent with each scaled vector. This will provide molecular dimer

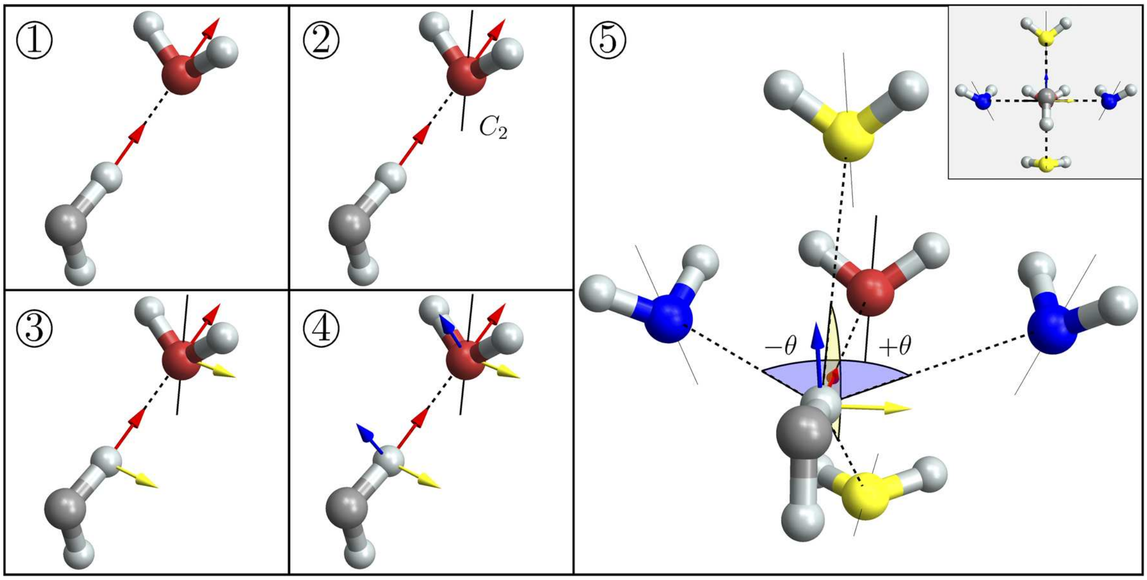

FIG. 2. Graphical depiction of steps 1–5 in the protocol used for generating four (of eight) non-equilibrium intermolecular angles for the water dimer. As described in the text, a local reference frame (*x* axis: blue solid arrow, *y* axis: yellow solid arrow, and *z* axis: red solid arrow) is defined with respect to the characteristic intermolecular interaction vector (black dashed line) between monomers *A* (red) and *B* (gray) as well as the principal axis on monomer *A* (black solid line). Preliminary geometries for the first four non-equilibrium intermolecular angles are then obtained by rotating *A* around the *x* and *y* axes on *B* by $\theta = \pm 30°$. For clarity, the inset of step 5 also provides a view down the *z* axis of the corresponding non-equilibrium geometries. To obtain preliminary geometries for the remaining four non-equilibrium intermolecular angles, this procedure is repeated after swapping the monomer labels. See Sec. II B for more details.

geometries along four separate intermolecular dissociation curves corresponding to each of the four non-equilibrium intermolecular angles.

STEP 8. Switch the *A* and *B* labels, and repeat steps 1–7. This will provide molecular dimer geometries along the intermolecular dissociation curves corresponding to each of the remaining four non-equilibrium intermolecular angles (for a total of eight non-equilibrium intermolecular angles). Note that the ion–π complexes in NENCI-2021 have only four unique non-equilibrium intermolecular angles due to the spherical symmetry of the ion; as such, step 8 is unnecessary and can be skipped for these molecular dimers.

STEP 9. For the equilibrium intermolecular angle, also scale the corresponding (optimized) interaction vector by factors of 0.7×, 0.8×, 0.9×, 0.95×, 1.05×, and 1.1×, and rigidly translate *A* consistent with each scaled vector. This will provide molecular dimer geometries along the intermolecular dissociation curve corresponding to the equilibrium intermolecular angle.

Here, we note in passing that the procedure outlined above for generating non-equilibrium intermolecular complexes is just one of a number of different methods for doing so. For instance, sampling dimer geometries from *ab initio* molecular dynamics (AIMD) simulations (or meta-dynamics[100] to enhance sampling away from the equilibrium/global minimum structure) is arguably one of the most straightforward alternatives. However, the procedure employed in this work was primarily chosen because it provides a systematic and well-defined way to sample the intermolecular PES (along the characteristic intermolecular interaction vector) for a large number of molecular dimers. In a follow-up manuscript, in this series (i.e., Paper II[91]), we will critically assess the performance of a large number of popular DFT and WFT methods on this database—here, the systematic structure of NENCI-2021 facilitates an analysis along the scaled intermolecular distance (i.e., the relevant correlation length for intermolecular interactions), which is not as straightforward with other sampling procedures.

C. Computational details

Intermolecular interaction energies (E_{int}) for each of the 7763 intermolecular complexes in NENCI-2021 were computed via

$$E_{\text{int}} = E_{AB} - E_A - E_B, \tag{1}$$

in which E_{AB} is the total energy of the dimer and E_A (E_B) is the total energy of monomer *A* (*B*). As mentioned above, all monomers were kept fixed at their optimized geometries, and the counterpoise correction of Boys and Bernardi[101] was applied to minimize basis set superposition error (BSSE).

Unless otherwise specified, Dunning's correlation consistent basis sets (with and without diffuse functions), namely, cc-pVXZ and aug-cc-pVXZ (with X = D, T, Q),[102–105] along with the frozen core (FC) approximation were used for all atoms except Li and Na. To provide a more accurate description of the core/valence electrons in the cation–π complexes, the cc-pwCVXZ[106] and aug-cc-pwCVXZ basis sets[106] were used for Li and Na in conjunction with the following modified FC approximation: $\text{Li}^+ = 1s^2$ (no core) and $\text{Na}^+ = [\text{He}]2s^2 2p^6$ ([He] core). All calculations employed the resolution-of-the-identity (RI) or density-fitting (DF) approximation during self-consistent field (SCF) calculations at the mean-field Hartree–Fock (HF) level and during post-HF calculations to account for electron correlation effects; the RI/DF approximation has been shown to introduce negligible errors when computing intermolecular interaction energies.[107,108] Whenever available, the corresponding JKFIT and RI auxiliary basis sets were used in conjunction with each primary (atomic orbital) basis set, i.e., cc-pVXZ-JKFIT/cc-pVXZ-RI[109,110] were used with cc-pVXZ, and aug-cc-pVXZ-JKFIT/aug-cc-pVXZ-RI[109,110] were used with aug-cc-pVXZ. For the cation–π complexes, the def2-aQZVPP-JKFIT/def2-aQZVPP-RI auxiliary basis sets[111–113] (which are some of the largest available auxiliary basis sets) were taken from the MOLPRO[114,115] basis set library and used in conjunction with cc-pwCVXZ and aug-cc-pwCVXZ for Li and Na. Throughout this work, we used the abbreviation aXZ to denote the following basis set usage: aug-cc-pVXZ (with aug-cc-pVXZ-JKFIT/aug-cc-pVXZ-RI) for {H, C, N, O, F, S, P, Cl, Br} and aug-cc-pwCVXZ (with def2-aQZVPP-JKFIT/def2-aQZVPP-RI) for {Li, Na}; we also use the abbreviation haXZ (i.e., heavy-aug-cc-pVXZ, also known as jul-cc-pVXZ[116]) to mean cc-pVXZ (with cc-pVXZ-JKFIT/cc-pVXZ-RI) for {H}, aug-cc-pVXZ (with aug-cc-pVXZ-JKFIT/aug-cc-pVXZ-RI) for {C, N, O, F, S, P, Cl, Br}, and aug-cc-pwCVXZ (with def2-aQZVPP-JKFIT/def2-aQZVPP-RI) for {Li, Na}.

For each molecular dimer and non-equilibrium intermolecular angle, the corresponding optimal intermolecular distance (1.0×) was obtained via a constrained minimization of E_{int} at the BSSE-corrected MP2/cc-pVTZ level [see Eq. (1)] following the procedure used to construct the S66 database;[57] for the molecular dimers not included in the original S66 database, the same procedure was also used to obtain the optimal intermolecular distance for the equilibrium intermolecular angle. In practice, this was accomplished by computing E_{int} for a series of dimer geometries in which monomer A (and/or B) was rigidly translated along the characteristic intermolecular interaction vector (see Sec. II B) and then locating the minimum value along the corresponding cubic spline interpolant. As such, each intermolecular complex in NENCI-2021 with a 1.0× intermolecular distance corresponds to the lowest-energy geometry (at the given intermolecular angle) at the BSSE-corrected MP2/cc-pVTZ level. For most non-covalent systems, optimization of the equilibrium molecular geometry at this level of theory yields structures that are close to the true minimum on the PES.[57] For the few molecular dimers that do not have a minimum using this method (e.g., ion–π complexes such as $\text{Li}^+/\text{Na}^+\cdots$ trifluorotriazine and $\text{F}^-/\text{Cl}^-\cdots$ benzene), the sum of the exchange, induction, and dispersion components of the SAPT2+/aDZ decomposition ($\varepsilon_{\text{EID}} = \varepsilon_{\text{Exch}} + \varepsilon_{\text{Ind}} + \varepsilon_{\text{Disp}}$, *vide infra*)—which have clear and distinct minima even for unbound ion–π complexes[48]—was minimized instead.

Benchmark E_{int} values in NENCI-2021 were obtained using Eq. (1) with all dimer (E_{AB}) and monomer (E_A and E_B) contributions computed using the "gold standard" CCSD(T) method extrapolated to the complete basis set (CBS) limit,[52,55] i.e.,

$$E^{\text{CCSD(T)/CBS}} \equiv E^{\text{MP2/CBS}} + \delta E^{\text{CCSD(T)/haTZ}}. \tag{2}$$

In this expression, the CBS-extrapolated MP2 total energy,

$$\begin{aligned} E^{\text{MP2/CBS}} &\equiv E^{\text{MP2/a(TQ)Z}} \\ &= E^{\text{HF/aQZ}} + E_{\text{corr}}^{\text{MP2/a(TQ)Z}}, \end{aligned} \tag{3}$$

was obtained using the two-point extrapolation procedure of Halkier *et al.*[117] on the MP2 correlation energy, namely,

$$E_{\text{corr}}^{\text{MP2/a(XY)Z}} = \frac{X^3 E_{\text{corr}}^{\text{MP2/aXZ}} - Y^3 E_{\text{corr}}^{\text{MP2/aYZ}}}{X^3 - Y^3} \tag{4}$$

with $X = 3$ (aTZ) and $Y = 4$ (aQZ). The so-called "delta" CCSD(T) correction,

$$\delta E^{\text{CCSD(T)/haTZ}} = E^{\text{CCSD(T)/haTZ}} - E^{\text{MP2/haTZ}}, \tag{5}$$

was computed using the haTZ basis set. The accuracy of this scheme for computing E_{int}—in particular for intermolecular complexes with particularly close contacts (i.e., 0.7× the equilibrium intermolecular separation)—is critically assessed in Sec. III C.

The energy decomposition analysis scheme (and classification of intermolecular binding motifs) provided in Sec. III B was based on calculations at the SAPT2+/aDZ level of theory,[88,118–121] the so-called "silver standard" of SAPT.[122]

All calculations in this work were performed using the `Psi4` (`v1.2`) software program.[123] During all HF calculations, the SCF convergence parameters were set to 1.0×10^{-8} in the total energy (`e_convergence = 1E-8`) and 1.0×10^{-8} in the root-mean-square DIIS error (`d_convergence = 1E-8`). For all CCSD(T) calculations, the CCSD convergence parameters were set to 1.0×10^{-6} in the total energy (`e_convergence = 1E-6`) and 1.0×10^{-5} in the residual of the t-amplitudes (`r_convergence = 1E-5`). All CCSD(T)/haTZ calculations (which constitute the vast majority of the computational effort needed to generate the benchmark E_{int} values in NENCI-2021) were completed in ≈ 1.5M core-hours using computational resources provided by *Cori Haswell* (Intel Xeon Processor E5-2698 v3 @ 2.30 GHz) and our research group cluster (Intel Xeon Platinum 8160 Processor @ 2.10 GHz).

D. Obtaining the NENCI-2021 database

A single `zip` file containing the Cartesian coordinates of the 7763 intermolecular complexes in NENCI-2021 (in `xyz` format) and a `csv` file containing all the CCSD(T)/CBS and SAPT energetic components (in kcal/mol) are provided as supplementary material. The properties of each monomer (i.e., charge, multiplicity, and the number of atoms), the corresponding benchmark E_{int} value, and the CCSD(T)/CBS and SAPT energetic components (in kcal/mol) can also be found in the comment line of each `xyz` file (see the included `README` file for additional details).

III. PROPERTIES OF THE NENCI-2021 DATABASE

A. Statistical analysis of intermolecular interaction energies and closest intermolecular contacts

A well-balanced database of intermolecular interactions should have a wide range of E_{int} values,[57] and this is indeed the case for NENCI-2021, as evidenced by the normalized E_{int} distributions provided in Fig. 3. With E_{int} values ranging from −38.5 to +186.8 kcal/mol, the benchmark intermolecular interaction energies in NENCI-2021 span 225.3 kcal/mol. In general, the most attractive (most negative) E_{int} values in NENCI-2021 correspond to charged intermolecular complexes that tend to be at or close to their equilibrium geometries. For instance, the single most attractive intermolecular complex in NENCI-2021 (with E_{int} = −38.5 kcal/mol) corresponds to the $\text{Li}^+\cdots$ benzene ion–π system at its equilibrium geometry (i.e., with Li^+ located above the center of the benzene ring; see Sec. II B). In fact, the top ten most attractive intermolecular interactions in NENCI-2021 correspond to the

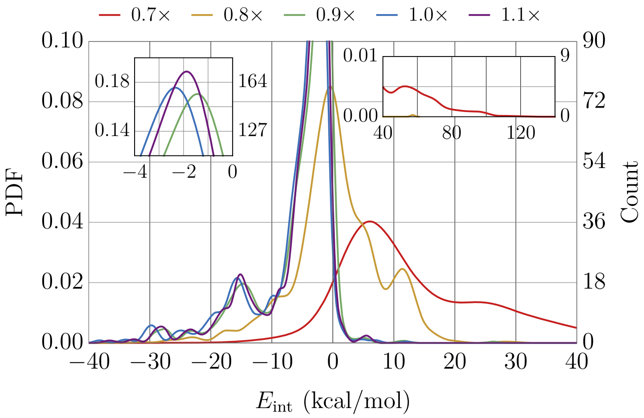

FIG. 3. Normalized probability density functions (PDFs) of the benchmark E_{int} values in the NENCI-2021 database as a function of the intermolecular distance (with the 0.95× and 1.05× scaled intermolecular distances omitted for clarity). The E_{int} values in NENCI-2021 range from −38.5 kcal/mol (most attractive) to +186.8 kcal/mol (most repulsive), with a mean (median) interaction energy of −1.06 kcal/mol (−2.39 kcal/mol). Insets display the peaks of the 0.9×, 1.0×, and 1.1× PDFs and the (positive E_{int}) tails of the 0.7× and 0.8× PDFs.

various $Li^+ \cdots \pi$ complexes at slightly different (but close to equilibrium) intermolecular distances and angles; these are followed by the ionic $H_2O \cdots HPO_4^{2-}$ hydrogen-bonded complexes (with a minimum $E_{\text{int}} = -34.6$ kcal/mol) and the $Na^+ \cdots \pi$ complexes (with a minimum $E_{\text{int}} = -28.2$ kcal/mol). In general, the most repulsive (most positive) E_{int} values in NENCI-2021 correspond to intermolecular complexes in which the monomers are separated by the shortest distance (0.7×) and rotated away from their equilibrium intermolecular angle, as both of these geometric perturbations lead to a rapid increase in the exponentially repulsive steric contribution to the interaction energy. For instance, the single most repulsive intermolecular complex in NENCI-2021 (with $E_{\text{int}} = +186.8$ kcal/mol) corresponds to the dimethyl sulfoxide (DMSO) dimer separated by 0.7× the equilibrium intermolecular distance and rotated to a non-equilibrium angle; in fact, this dimer has the closest intermolecular contact in the entire database (with $d_{\text{H}\cdots\text{H}} = 0.81$ Å, *vide infra*). Other substantially repulsive intermolecular complexes in NENCI-2021 include the $Na^+ \cdots$ trifluorotriazine ion–π system (with a maximum $E_{\text{int}} = +118.8$ kcal/mol) and another DMSO dimer (with $E_{\text{int}} = +112.4$ kcal/mol), both of which were characterized by a 0.7× intermolecular separation and a non-equilibrium intermolecular angle.

The mean and median E_{int} values in NENCI-2021 are −1.1 kcal/mol and −2.4 kcal/mol, respectively, which correspond to typical interaction energies found in weakly bound molecular dimers. These statistical measures are primarily governed by the (relatively) large number of intermolecular complexes in NENCI-2021 that contain monomers in non-equilibrium (angular) orientations. Such geometric perturbations tend to nullify the energetic stabilization provided by *directional* intermolecular binding motifs (e.g., single and double hydrogen bonds, and dipole–dipole interactions) and often result in complexes with weakly attractive E_{int} values. Broken down by the scaled intermolecular distance, the mean (median) E_{int} values are as follows: +21.8 (+11.7) kcal/mol for 0.7×, +0.5 (+0.2) kcal/mol for 0.8×, −5.2 (−2.9) kcal/mol for 0.9×, −6.0 (−3.5) kcal/mol for 0.95×, −6.1 (−3.5) kcal/mol for 1.0×, −5.9 (−3.4) kcal/mol for 1.05×, and −5.5 (−3.1) kcal/mol for 1.10×. In total, NENCI-2021 contains 6020 attractive ($E_{\text{int}} < 0$) and 1743 repulsive ($E_{\text{int}} > 0$) intermolecular complexes, and the crossover from attractive to repulsive E_{int} values typically occurs around 0.8× the equilibrium intermolecular distance. As one might expect, the proportion of attractive intermolecular interactions in NENCI-2021 quickly diminishes as the distance between monomers decreases; broken down again by the scaled intermolecular distance, we find that the percentage of attractive (repulsive) E_{int} values are as follows: 3.4% (96.6%) for 0.7×, 47.8% (52.2%) for 0.8×, 97.6% (2.4%) for 0.9×, 98.3% (1.7%) for 0.95×, 98.3% (1.7%) for 1.0×, 98.2% (1.8%) for 1.05×, and 98.0% (2.0%) for 1.1×. Quite interestingly, there are still a number ($N = 38$) of attractive intermolecular complexes at the 0.70× scaled intermolecular distance, which generally correspond to strongly favorable dimers such as the $Li^+ \cdots \pi$ complexes discussed above. In the same breath, there are also quite a few ($N = 19$) repulsive complexes at the equilibrium (1.0×) distance—some of which even occur at the corresponding equilibrium angle, e.g., the cation–π and anion–π complexes involving trifluorotriazine and benzene, respectively.

A well-balanced database of intermolecular interactions should also sample a wide range of intermolecular atom-pair distances (i.e., interatomic distances between the atoms on molecule A and the atoms on molecule B). Again, this is indeed the case for NENCI-2021, and is demonstrated by the series of normalized probability density functions (PDFs) in Fig. 4, which quantify a representative set of atom-pair distances (i.e., $O \cdots H$, $N \cdots H$, $H \cdots H$, and $C \cdots H$) as a function of intermolecular separation. In Fig. 4, we chose to focus on the $O \cdots H$, $N \cdots H$, $H \cdots H$, and $C \cdots H$ intermolecular atom-pair distances as the first two are representative of hydrogen-bonded systems and the last two are the relevant interatomic distances for non-bonded complexes in general. Since NENCI-2021 was designed with a particular emphasis on close intermolecular contacts, we focus our discussion on the short-distance sectors in these PDFs. As discussed above in the Introduction, such close intermolecular contacts are important in a number of applications[18–21,49] and pose significant difficulty for both WFT and DFT methods (see Paper II[91] in this series), as both strongly attractive and strongly repulsive intermolecular forces must be accurately described to obtain a quantitatively correct E_{int} value. As the intermolecular distance is reduced from 1.1× to 0.7×, the complexes in NENCI-2021 sample increasingly closer interatomic distances and begin to more appreciably populate the region inside the corresponding vdW envelope. In other words, a number of intermolecular atom-pair distances (R_{AB}) are less than the sum of the corresponding vdW radii, i.e., $R_{AB} < R_{AB}^{\text{vdW}} \equiv R_A^{\text{vdW}} + R_B^{\text{vdW}}$. Plotted as vertical black dotted lines in Fig. 4, these R_{AB}^{vdW} values were computed using the vdW radii provided by Bondi[124] for {C, N, O} and the revised value obtained by Rowland and Taylor[125] for {H}, and take on the following values: 2.61 Å ($O \cdots H$), 2.64 Å ($N \cdots H$), 2.18 Å ($H \cdots H$), and 2.79 Å ($C \cdots H$). As one would expect, these values are always smaller than the mean closest contact distances for the 141 equilibrium intermolecular complexes in NENCI-2021, i.e., 2.68 Å ($O \cdots H$), 3.09 Å ($N \cdots H$), 3.01 Å ($H \cdots H$), and 3.17 Å ($C \cdots H$); for comparative purposes, these values are plotted as vertical blue solid lines in Fig. 4. Broken down by scaled intermolecular

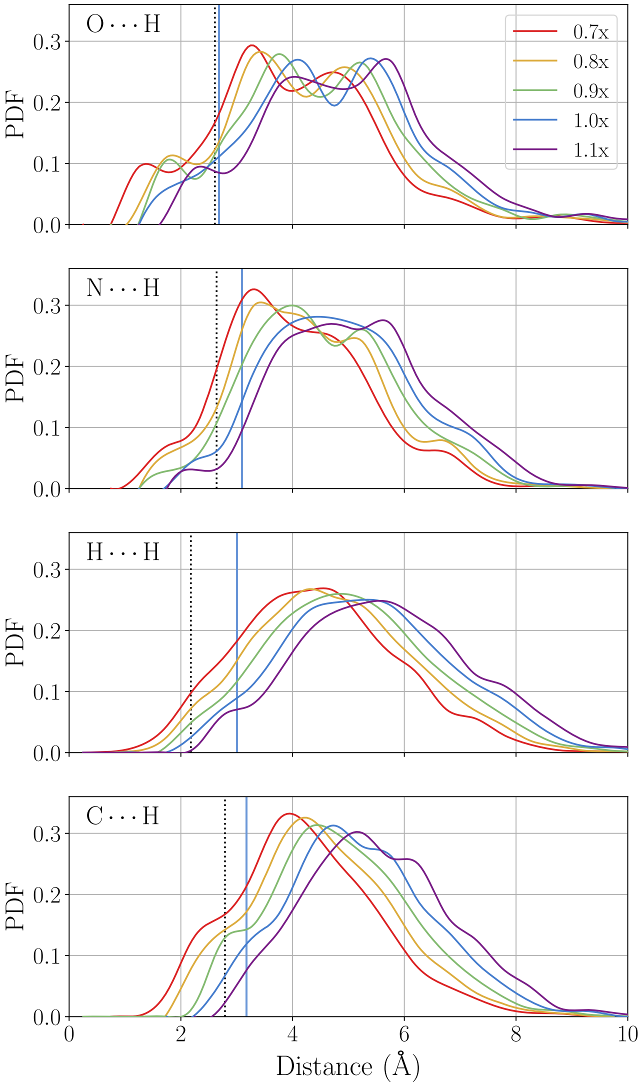

FIG. 4. Normalized probability density functions (PDFs) of the O···H, N···H, H···H, and C···H intermolecular atom-pair distances in the NENCI-2021 database as a function of the scaled intermolecular distance (with the 0.95× and 1.05× scaled intermolecular distances omitted for clarity). For reference, vertical blue solid lines indicate the mean closest intermolecular contact distances in the 141 equilibrium complexes in NENCI-2021, while the vertical black dotted lines indicate the sum of the atomic vdW radii[124,125] corresponding to each atom pair. See the main text for more details.

distance, the percentage of O···H (N···H) intermolecular atom-pair distances with $R_{AB} < R_{AB}^{\mathrm{vdW}}$ is 17.0% (10.8%) for 0.7×, 13.5% (7.4%) for 0.8×, 10.7% (4.4%) for 0.9×, 9.1% (3.4%) for 0.95×, 7.8% (2.8%) for 1.0×, 6.9% (2.4%) for 1.05×, and 6.1% (1.9%) for 1.1×. Applying the same analysis to the H···H (C···H) atom-pair distances yields the following: 4.5% (13.2%) for 0.7×, 2.4% (8.6%) for 0.8×, 1.2% (4.4%) for 0.9×, 0.7% (2.5%) for 0.95×, 0.2% (1.2%) for 1.0×, 0.1% (0.5%) for 1.05×, and 0.1% (0.2%) for 1.1×. The closest contacts in NENCI-2021 occur in complexes in which the intermolecular distance has been scaled to 0.7× and the monomers have been rotated such that the atoms (on each monomer) are forced into close proximity. For reference, the shortest intermolecular atom-pair distances in NENCI-2021 are significantly shorter than R_{AB}^{vdW}, and were found to be 0.82 Å for O···H (DMSO dimer, E_{int} = +186.8 kcal/mol, 31% of $R_{\mathrm{OH}}^{\mathrm{vdW}}$), 1.30 Å for N···H (uracil···neopentane, E_{int} = +80.5 kcal/mol, 49%), 0.81 Å for H···H (peptide···pentane, E_{int} = +72.3 kcal/mol, 37%), and 1.22 Å for C···H (benzene···AcOH, E_{int} = +76.4 kcal/mol, 44%). While it is not surprising to find the most repulsive intermolecular complex in NENCI-2021 (i.e., the DMSO dimer) listed among the closest contacts, the other examples are far from the most positive end of the E_{int} spectrum and reflect the wide range of attractive and repulsive intermolecular forces sampled in this database.

B. Energy decomposition analysis of the intermolecular binding motifs

A well-balanced database of intermolecular interactions should also sample a wide variety of different binding motifs. Here, we would again argue that this is the case for NENCI-2021, and demonstrate this point by the extensively populated ternary diagrams depicted in Fig. 5. Introduced by Singh *et al.*[126] in the late 2000s, these ternary diagrams were constructed using a SAPT decomposition of E_{int} into the following four components for each intermolecular complex in NENCI-2021: $\varepsilon_{\mathrm{Elst}}$ (electrostatics, Elst), $\varepsilon_{\mathrm{Exch}}$ (exchange, Exch), $\varepsilon_{\mathrm{Ind}}$ (induction, Ind), and $\varepsilon_{\mathrm{Disp}}$ (dispersion, Disp), i.e., $E_{\mathrm{int}} \approx \varepsilon_{\mathrm{SAPT}} = \varepsilon_{\mathrm{Elst}} + \varepsilon_{\mathrm{Exch}} + \varepsilon_{\mathrm{Ind}} + \varepsilon_{\mathrm{Disp}}$. In particular, we performed this decomposition at the SAPT2+/aDZ level of theory,[88,118–121] the so-called "silver standard" of SAPT,[122] which has been shown to have an overall mean absolute error (MAE) of 0.30 kcal/mol across the S22,[54] HBC6,[127] NBC10,[128–131] and HSG[67] databases.[55] Unlike the "bronze standard" sSAPT0/jun-cc-pVDZ,[122] which can underestimate the dispersion component in anion–π complexes by more than 100%,[132] the more sophisticated SAPT2+/aDZ method employed herein is expected to more accurately describe $\varepsilon_{\mathrm{Disp}}$ in the 700 anion–π complexes present in NENCI-2021. As such, this SAPT level should be well-suited to provide a physically sound and semi-quantitative characterization of the binding motifs included in NENCI-2021.

In previously constructed databases of non-covalent interactions (e.g., S66[57] and S101[49]), each intermolecular complex was typically classified into one of three categories based on whether $E_{\mathrm{int}} \approx \varepsilon_{\mathrm{SAPT}}$ was dominated by the $\varepsilon_{\mathrm{Elst}}$ component (Elst-bound), the $\varepsilon_{\mathrm{Disp}}$ component (Disp-bound), or a mixture thereof (Mix-bound). Since the $\varepsilon_{\mathrm{Ind}}$ component tended to be small in these complexes, the analogous and fourth Ind-bound category was deemed to be largely unnecessary. With the addition of 1400 ion–π complexes (in particular, the 700 cation–π systems), the scope of the SAPT decomposition analysis is substantially wider in NENCI-2021 and now encompasses the Ind-bound regime.[48] As such, we propose a natural extension of the traditional three-category classification scheme made popular by Řezáč *et al.*[57] and Burns *et al.*[53,133] to include the Ind-bound category. To do so, we construct a three-dimensional feature space defined by the $\varepsilon_{\mathrm{Disp}}/\varepsilon_{\mathrm{Elst}}$, $\varepsilon_{\mathrm{Ind}}/\varepsilon_{\mathrm{Disp}}$, and $\varepsilon_{\mathrm{Elst}}/\varepsilon_{\mathrm{Ind}}$ ratios as follows:

STEP 1. To start, a single dimension of the feature space is chosen as the basis for constructing an initial sub-classification scheme. Although this choice is arbitrary, we will start with the $\varepsilon_{\mathrm{Disp}}/\varepsilon_{\mathrm{Elst}}$

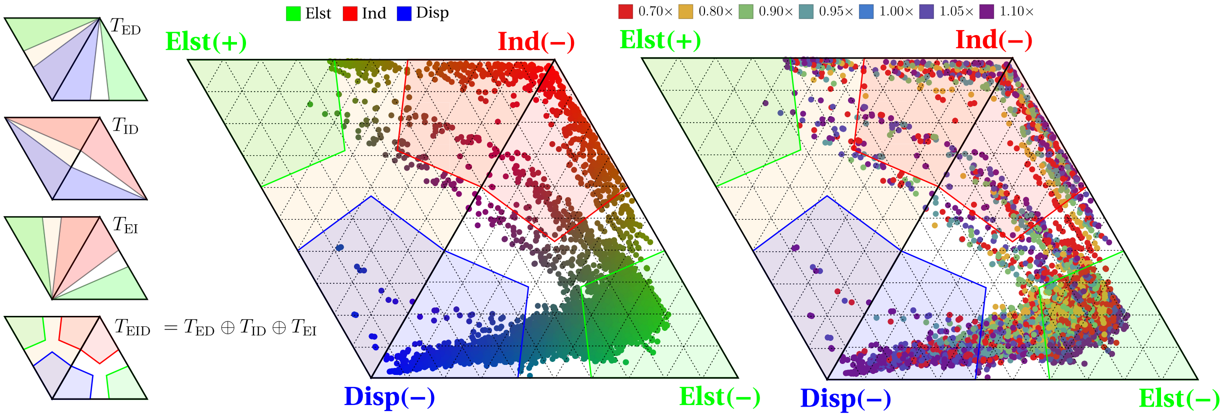

FIG. 5. (*left*) Geometric depiction of the extended four-category classification scheme (based on a SAPT decomposition of E_{int}) used to classify each intermolecular complex in NENCI-2021 as follows: Elst-bound (E, green), Ind-bound (I, red), Disp-bound (D, blue), or Mix-bound. As described in the main text, this classification scheme can be represented by a fused ternary diagram (T_{EID}), which has been colored according to the following rule: each intermolecular complex that has been assigned the same category in any two of the three sub-classification schemes (T_{ED}, T_{ID}, and T_{EI}) retains that color in T_{EID}; otherwise, the complex is classified as Mix-bound (white). (*middle/right*) Ternary diagrams depicting the breakdown of the SAPT2+/aDZ intermolecular interaction energies of each complex in NENCI-2021 according to the contributions from electrostatics (ε_{Elst}), induction (ε_{Ind}), and dispersion (ε_{Disp}). Since ε_{Elst} can be positive [Elst(+)] or negative [Elst(−)], these plots are comprised of two ternary diagrams [one for Elst(+) and one for Elst(−)] that have been fused together. In these ternary diagrams, the shaded polygons are used to reflect the four-category classification scheme described above, i.e., Elst-bound (green), Ind-bound (red), and Disp-bound (blue); complexes that are not located in any one of these regions are Mix-bound. In the *middle* panel, each point has been colored using an RGB scheme with values given by $\{|\varepsilon_{Ind}|/(|\varepsilon_{Elst}| + |\varepsilon_{Ind}| + |\varepsilon_{Disp}|),\ |\varepsilon_{Elst}|/(|\varepsilon_{Elst}| + |\varepsilon_{Ind}| + |\varepsilon_{Disp}|),\ |\varepsilon_{Disp}|/(|\varepsilon_{Elst}| + |\varepsilon_{Ind}| + |\varepsilon_{Disp}|)\}$. In the *right* panel, each point is colored according to the scaled intermolecular distance.

ratio, as this selection is tantamount to constructing the aforementioned three-category classification scheme (i.e., Elst-bound, Disp-bound, or Mix-bound). For illustrative purposes, a ternary diagram (T_{ED}) depicting this initial sub-classification scheme is plotted in the left panel of Fig. 5.

STEP 2. Intermolecular complexes with $|\varepsilon_{Disp}/\varepsilon_{Elst}| > \eta$ are sub-classified as Disp-bound (blue shaded regions in T_{ED}), while intermolecular complexes with $|\varepsilon_{Elst}/\varepsilon_{Disp}| > \eta$ are sub-classified as Elst-bound (green shaded regions in T_{ED}). If one stopped at this point, set $\eta = 2$, and classified all other cases as Mix-bound, this initial sub-classification scheme (based on the single $\varepsilon_{Disp}/\varepsilon_{Elst}$ feature) would be equivalent to the three-category classification scheme described above. Since the value of η is somewhat arbitrary, we have chosen to employ a slightly smaller value ($\eta = 3/2$) in the classification scheme introduced in this work; with this choice for η, fewer intermolecular complexes will be classified as Mix-bound (*vide infra*).

STEP 3. To go beyond this three-category classification scheme, step 2 is repeated for the two remaining dimensions of the feature space. Selection of the $\varepsilon_{Ind}/\varepsilon_{Disp}$ feature generates the T_{ID} ternary diagram in Fig. 5 and the analogous sub-classification of intermolecular complexes as follows: Disp-bound (if $\varepsilon_{Disp}/\varepsilon_{Ind} > \eta$; blue shaded regions in T_{ID}) or Ind-bound (if $\varepsilon_{Ind}/\varepsilon_{Disp} > \eta$; red shaded regions). Similarly, the $\varepsilon_{Elst}/\varepsilon_{Ind}$ feature yields the final required sub-classification scheme: Elst-bound (if $|\varepsilon_{Elst}/\varepsilon_{Ind}| > \eta$; green shaded regions in T_{EI}) or Ind-bound (if $|\varepsilon_{Ind}/\varepsilon_{Elst}| > \eta$; red shaded regions). Here, we note in passing that the absolute value (magnitude) must be used for all sub-classifications based on ε_{Elst} as the sign of the Elst component can be positive or negative.

STEP 4. To arrive at our extended (i.e., four-category) classification scheme, each intermolecular complex that has been sub-classified (in step 2 and step 3) with the same label twice retains that label; otherwise, the intermolecular complex is classified as Mix-bound. This final classification scheme is graphically depicted in the colored T_{EID} ternary diagram in Fig. 5, which is assembled as an "outer sum" over the colored ternary diagrams corresponding to the sub-classification schemes, i.e., $T_{EID} = T_{ED} \oplus T_{ID} \oplus T_{EI}$, in which the colors of T_{EID} are determined according to the rules described above.

Based on this extended four-category classification scheme, the 141 equilibrium intermolecular complexes in NENCI-2021 are comprised of 54 (38.3%) Elst-bound, 23 (16.3%) Ind-bound, 31 (22.0%) Disp-bound, and 33 (23.4%) Mix-bound dimers. When including all non-equilibrium intermolecular distances and angles, the entire NENCI-2021 database contains 3499 (45.1%) Elst-bound, 700 (9.0%) Ind-bound, 1372 (17.7%) Disp-bound, and 2192 (28.2%) Mix-bound intermolecular complexes. Here, we note in passing that this observed decrease in the percentage of Ind-bound complexes is partially due to the inclusion of five (instead of nine) intermolecular angles for each ion–π complex due to symmetry considerations (see Sec. II B). As such, the intermolecular complexes in NENCI-2021 largely span the entire ternary diagram in Fig. 5 and therefore contain a diverse array of binding motifs; as result, we hope that NENCI-2021 will be used to critically examine (and potentially improve) the performance of theoretical models when faced with the challenge of simultaneously describing diverse NCI types on the same footing [i.e., point (ii) in the Introduction].

Here, we note that the apparent bias toward Elst-bound complexes in NENCI-2021 is an unavoidable consequence of sampling short-range intermolecular separations; at such distances, there is often a substantial amount of orbital/density overlap between monomers, and charge penetration effects[5,48,61,87,99,134] (in $\varepsilon_{\text{Elst}}$) tend to be the dominant contribution (over ε_{Ind} and $\varepsilon_{\text{Disp}}$) to $\varepsilon_{\text{SAPT}}$. For instance, a significant majority (73.9%) of the intermolecular complexes at 0.7× are classified as Elst-bound, while approximately half that (38.3%) of the 141 equilibrium dimers share this label. This increase in the relative number of Elst-bound complexes at shorter intermolecular separations is clearly reflected in the ternary diagram in the right panel of Fig. 5 and the percentage of Elst-bound complexes when broken down by the scaled intermolecular distance, i.e., 73.9% (0.7×), 48.9% (0.8×), 40.5% (0.9×), 38.1% (0.95×), 37.3% (1.0×), 38.0% (1.05×), and 38.9% (1.1×). In general, many complexes that are Disp-bound at larger intermolecular distances become Elst-bound or Mix-bound at reduced separations where short-range effects (e.g., charge penetration) become more significant. On the other hand, the Ind-bound complexes (which are primarily comprised of cation–π interactions) tend to remain Ind-bound even at reduced intermolecular separations since charge penetration effects are substantially reduced when one of the monomers is a monovalent cation (e.g., Li^+ or Na^+).[48] For reference, the respective percentages of Ind-bound, Disp-bound, or Mix-bound complexes as a function of the scaled intermolecular distance are as follows: 6.0%, 0.0%, and 20.2% for 0.7×; 7.6%, 0.0%, and 43.6% for 0.8×; 8.6%, 15.2%, and 35.7% for 0.9×; 10.0%, 22.8%, and 29.0% for 0.95×; 10.3%, 27.8%, and 24.6% for 1.0×; 10.3%, 28.7%, and 23.1% for 1.05×; and 10.5%, 29.2%, and 21.5% for 1.1×.

Before moving on to consider the error/uncertainty in the E_{int} values in NENCI-2021, we note in passing that the positive electrostatics [Elst(+)] region of the ternary diagram in Fig. 5 does not appear to be well sampled. However, NENCI-2021 does contain a non-negligible (413) number of intermolecular complexes with $\varepsilon_{\text{Elst}} > 0$. As mentioned above, such complexes are primarily found among the cation–π complexes, where the degree of orbital overlap in the dimer (and hence the energetic stabilization due to charge penetration effects) is largely suppressed;[48] hence, intermolecular complexes with repulsive $\varepsilon_{\text{Elst}}$ values are quite rare and may be adequately accounted for in NENCI-2021.

C. Error analysis and critical assessment of the benchmark intermolecular interaction energies

In addition to being extensive in size and scope, we would also argue that a well-balanced database of intermolecular interactions should contain a reliable estimate of the error/uncertainty present in the computed E_{int} values. For *ab initio* WFT methods, the two primary sources of error when computing E_{int} are as follows: (i) incompleteness in the one-particle basis set [i.e., basis set incompleteness error (BSIE)] and (ii) the approximate treatment of the electron correlation energy (E_{corr}). Since the mean-field HF contribution to E_{int} converges quickly with respect to the underlying basis set,[135] we expect that the BSIE at the $E_{\text{int}}^{\text{HF/aQZ}}$ level will be negligible when compared to the BSIE in the post-HF correlation energy contributions in Eq. (2). As depicted in Eq. (3), the BSIE in the MP2 correlation energy is largely mitigated using the two-point extrapolation scheme[117] for approximating the MP2/CBS limit provided in Eq. (4). Although the $\delta E^{\text{CCSD(T)}}$ correction tends to converge quickly with respect to the basis set (in part, due to the relatively small size of the correction[75]),[128,136–138] the BSIE in this term is generally the largest remaining source of error for extrapolation schemes such as that outlined in Eqs. (2)–(5).[55,139] To mitigate this error (and still remain computationally feasible when generating such a large number of intermolecular interaction energies), this contribution was computed using an augmented Dunning-style triple-ζ (haTZ) basis set[102–105] in NENCI-2021 [cf. Eq. (5)].

As such, we will primarily focus on the remaining BSIE in the $\delta E^{\text{CCSD(T)/haTZ}}$ contribution to E_{int} when critically assessing the accuracy of the intermolecular interaction energies in NENCI-2021. To do so, we will compare our E_{int} values against two different references. As a first reference value, we computed the $\delta E^{\text{CCSD(T)}}$ correction in Eq. (5) using a larger (and substantially more expensive) augmented quadruple-ζ (aQZ) basis set, i.e.,

$$\begin{aligned} E^{\text{REF1}} &= E^{\text{MP2/CBS}} + \delta E^{\text{CCSD(T)/aQZ}} \\ &= E^{\text{HF/aQZ}} + E_{\text{corr}}^{\text{MP2/a(TQ)Z}} + \delta E^{\text{CCSD(T)/aQZ}}, \end{aligned} \tag{6}$$

in which $E^{\text{MP2/CBS}}$ was computed using Eqs. (3) and (4). As a second and alternative reference, we simply replaced the $\delta E^{\text{CCSD(T)/haTZ}}$ correction with a direct two-point extrapolation[117] of $E^{\text{CCSD(T)}}$ using the aTZ and aQZ basis sets, i.e.,

$$\begin{aligned} E^{\text{REF2}} &= E^{\text{CCSD(T)/a(TQ)Z}} \\ &= E^{\text{HF/aQZ}} + E_{\text{corr}}^{\text{CCSD(T)/a(TQ)Z}}. \end{aligned} \tag{7}$$

By including CCSD(T) calculations in the much larger aQZ basis set, both of these reference values directly probe the BSIE in the CCSD(T) contribution and are expected to be more reliable than the E_{int} values in the NENCI-2021 database.

The error of the CCSD(T)/CBS scheme outlined in Eqs. (2)–(5) with respect to both E^{REF1} and E^{REF2} is shown in Fig. 6 for a select subset of intermolecular complexes in NENCI-2021. Plotted as a function of the scaled intermolecular distance (at the equilibrium angle, unless otherwise noted), this subset of intermolecular complexes was chosen to cover the wide array of binding motifs in NENCI-2021 and includes examples of Elst-, Ind-, Disp-, and Mix-bound systems, i.e., single ($H_2O\cdots H_2O$, $MeNH_2\cdots MeNH_2$) and double (AcOH$\cdots$AcOH) hydrogen bonds, dipole–dipole (MeF$\cdots$MeF), π–π stacking (BZ$\cdots$BZ PD), CH-π (BZ $\cdots$ BZ TS), and cation–π ($Na^+\cdots$BZ) and anion–π ($F^-\cdots$BZ) interactions. As seen in Fig. 6, the E_{int} values in NENCI-2021 are generally within ±0.1 kcal/mol of both E^{REF1} and E^{REF2}, and the errors with respect to these references tend to increase in magnitude at reduced intermolecular distances. The worst-case scenarios among this subset include the acetic acid dimer (AcOH$\cdots$ AcOH, double hydrogen-bonded) and the C_{2h} parallel-displaced (PD) benzene dimer (BZ $\cdots$ BZ PD, π–π stacking), with errors in both steadily increasing in magnitude as the intermolecular separation is decreased; at 0.7×, we report errors of +0.19 kcal/mol (AcOH$\cdots$ AcOH) and −0.15 kcal/mol (BZ $\cdots$ BZ PD) with respect to E^{REF1} (+0.30 and −0.27 kcal/mol when compared to E^{REF2}, *vide infra*). In these cases, the increased error is most likely due to the relatively larger amount of orbital overlap between these monomers at reduced intermolecular separations, where the interplay between short-range intermolecular interactions (i.e., charge penetration, Pauli repulsion, many-body exchange–correlation effects, etc.) becomes increasingly more challenging to describe in an accurate and reliable fashion.

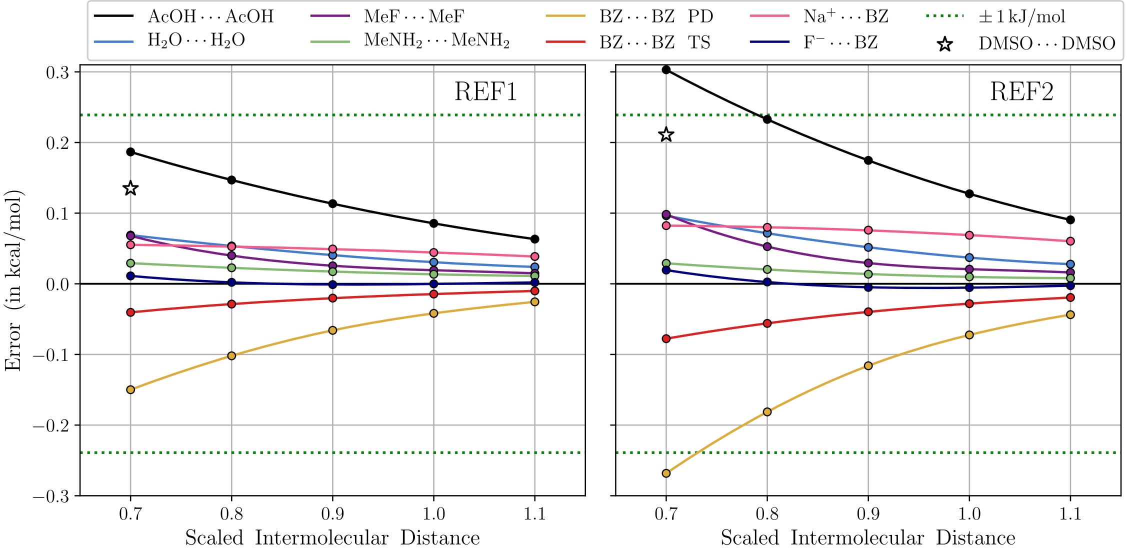

FIG. 6. Errors (in kcal/mol) in the NENCI-2021 E_{int} values [computed using the $E^{\text{CCSD(T)/CBS}}$ extrapolation scheme in Eqs. (2)–(5)] with respect to E^{REF1} [Eq. (6); *left*] and E^{REF2} [Eq. (7); *right*] for a representative set of intermolecular complexes. Plotted as a function of the scaled intermolecular distance (with the 0.95× and 1.05× distances omitted for clarity), all intermolecular complexes (with the exception of DMSO· · ·DMSO, black stars) were kept at their equilibrium angle. Errors with respect to E^{REF1} and E^{REF2} were computed as $E^{\text{CCSD(T)/CBS}} - E^{\text{REF1}}$ and $E^{\text{CCSD(T)/CBS}} - E^{\text{REF2}}$, respectively. Based on this error profile (see the main text), the errors in the NENCI-2021 E_{int} values are ±0.1 kcal/mol on average but can be as large as ±0.2–0.3 kcal/mol (i.e., ±1 kJ/mol, green dashed lines) for some complexes at reduced (i.e., 0.7× and 0.8×) intermolecular separations.

This trend is also reflected in the error profiles corresponding to the two different BZ ··· BZ dimers in Fig. 6, where one can see that the error in the PD dimer (more orbital overlap) is noticeably larger in magnitude than the error in the C_{2v} T-shaped (TS) dimer (less orbital overlap) at all intermolecular separations. In the same breath, we also note that the error with respect to E^{REF1} (or E^{REF2}) is non-trivial in general, and does not necessarily follow a direct/straightforward correlation with closest intermolecular contacts and/or the sign/magnitude of E_{int}. For example, the errors for AcOH··· AcOH (E_{int} = +12.1 kcal/mol) and BZ ··· BZ PD (E_{int} = +51.2 kcal/mol) at 0.7× are both larger than that found in the intermolecular complex with the largest (most repulsive) E_{int} value and closest O···H distance in NENCI-2021—a non-equilibrium configuration of DMSO ··· DMSO with E_{int} = +186.8 kcal/mol (whose errors with respect to E^{REF1} and E^{REF2} are depicted by stars in Fig. 6).

From this analysis, we believe that the errors in the NENCI-2021 E_{int} values are mostly within ±0.1 kcal/mol, but can be as large as 0.2–0.3 kcal/mol (i.e., ≈ 1 kJ/mol) for certain systems at reduced intermolecular separations. Here, we note in passing that the $\delta E^{\text{CCSD(T)/haTZ}}$ correction used in NENCI-2021 provides a significant improvement over $\delta E^{\text{CCSD(T)/aDZ}}$ and yields nearly identical E_{int} values when compared to the more expensive $\delta E^{\text{CCSD(T)/aTZ}}$ approach; this is shown in Fig. S1 and again emphasizes the need for triple-ζ basis sets when employing the $\delta E^{\text{CCSD(T)}}$ correction scheme.[55] When considering the largest errors in Fig. 6, i.e., AcOH···AcOH and BZ···BZ PD, one can see that the errors with respect to E^{REF1} and E^{REF2} differ by ≈ 0.1 kcal/mol; as such, the estimated *average* error in NENCI-2021 (±0.1 kcal/mol) is comparable to the difference between using E^{REF1} or E^{REF2} as the reference for E_{int}. Generally speaking, it is not clear which of these two quantities supplies the more accurate reference for E_{int}; however, it has been pointed out by Marshall *et al.*[55] that the $\delta E^{\text{CCSD(T)}}$ correction does not converge monotonically toward the CBS limit, which implies that E^{REF1} might in fact be a slightly better reference value than E^{REF2}.

As mentioned above, the other primary source of error when computing E_{int} using approximate *ab initio* WFT methods is the necessarily incomplete treatment of the electron correlation energy; while post-CCSD(T) corrections tend to be small for *equilibrium* intermolecular interaction energies (i.e., < 0.1 kcal/mol),[52] whether or not such corrections become more substantial at reduced intermolecular separations still remains unanswered. With increasingly unfavorable scaling with both system and basis set size, such post-CCSD(T) calculations [i.e., CCSDT, CCSDT(Q), CCSDTQ, etc.] are computationally prohibitive and could have only been performed on the following: (i) the smaller/smallest systems in NENCI-2021, but with sufficiently large basis sets (of at least triple-ζ or quadruple-ζ quality) or (ii) the larger/largest systems in NENCI-2021, but with reduced and insufficiently large basis sets (i.e., double-ζ at best). Since neither of these approaches would have provided an accurate and reliable estimate of the post-CCSD(T) contributions to E_{int} for the wide range of intermolecular complexes in NENCI-2021,[140] we chose to focus our efforts above on critically assessing the CCSD(T)/CBS scheme outlined in Eqs. (2)–(5) based on a quantitative estimate of the remaining BSIE at the CCSD(T) level. Since an accurate and reliable prediction of E_{int} for intermolecular complexes in the repulsive wall (i.e., inside the vdW envelope) poses a substantive challenge to state-of-the-art DFT and WFT methods (see Paper II[91] in this series), further benchmarking of the standard CCSD(T)/CBS approach [possibly via stochastic CC[141–143] or full configuration interaction (FCI)[144–146] methods] in this regime is an open challenge for the community and will be of critical importance for the development of next-generation DFT functionals and ML-based intra-/inter-molecular interaction potentials.

IV. CONCLUSIONS AND FUTURE DIRECTIONS

In this work, we present NENCI-2021: a large and comprehensive database of ∼8000 benchmark Non-Equilibirum Non-Covalent Interaction energies for a diverse selection of intermolecular complexes of biological and chemical relevance with a particular emphasis on close intermolecular contacts. Designed to address the growing need for extensive high-quality quantum mechanical data in the chemical sciences, NENCI-2021 starts with the 101 molecular dimers in the widely used S66, S66x8, S66a8, and S101x7 databases[49,57,59] and extends the scope of these popular works in two directions. For one, NENCI-2021 includes 40 cation–π and anion–π complexes, a fundamentally important class of NCIs that are found throughout nature and among the strongest NCIs known. Second, NENCI-2021 systematically samples both equilibrium and non-equilibrium configurations on all 141 intermolecular PESs by *simultaneously* varying the intermolecular distance (from 0.7× to 1.1× the equilibrium separation) and intermolecular angle (including either five or nine angles for each distance depending on symmetry considerations). As such, a wide range of intermolecular atom-pair distances are present in NENCI-2021, including a large number of close intermolecular contacts with atom pairs located inside their respective vdW envelope; these intermolecular complexes probe a number of different short-ranged NCIs (e.g., charge transfer and penetration, Pauli repulsion, and many-body exchange–correlation effects), which are observed in many important chemical and biological systems, and pose an enormous challenge for molecular modeling. Computed at the CCSD(T)/CBS level of theory, the 7763 benchmark E_{int} values in NENCI-2021 range from −38.5 kcal/mol (most attractive) to +186.8 kcal/mol (most repulsive), with a total span of 225.3 kcal/mol and a mean (median) E_{int} value of −1.06 kcal/mol (−2.39 kcal/mol). Of these 7763 interaction energies, ≈5% have E_{int} values greater than 20 kcal/mol, and would therefore be considered statistically negligible in most physical applications (i.e., $E_{\text{int}} \gg k_B T$). While such systems have been included in NENCI-2021 for completeness, we recommend that potential users of NENCI-2021 exert caution when employing these data points when testing/training approximate methods, as their inclusion could skew the error metric and/or overall statistics. A detailed SAPT-based analysis was used to confirm the diverse and comprehensive nature of the intermolecular binding motifs present in NENCI-2021, which includes a significant number of primarily induction-bound dimers and now spans all regions of the SAPT ternary diagram; this warranted a new four-category classification scheme that includes complexes primarily bound by electrostatics (3499), induction (700), dispersion (1372), or mixtures thereof (2192). Finally, a critical error analysis was performed on a representative set of intermolecular complexes, from which we estimate that the E_{int} values in NENCI-2021 have a mean error of ±0.1 kcal/mol and a maximum error of ±0.2–0.3 kcal/mol for the most challenging cases.

For all these reasons, we believe that the NENCI-2021 database is timely and well-suited for testing, training, and developing next-generation force fields, DFT and WFT methods, and ML-based potentials. In this regard, NENCI-2021 can be used for a variety of purposes. For one, NENCI-2021 could be employed as a single database and used in its entirety. Alternatively, NENCI-2021 can be split into multiple different training and testing data sets—each containing a diverse sample of intermolecular binding motifs—and used for cross-validation studies and statistical error assessment. When used for such purposes, we note in passing that strong correlations will likely exist between different points on a given intermolecular PES; as such, we caution against separating such points between training and testing datasets to avoid issues associated with overfitting. For ML applications, we expect NENCI-2021 to be useful for training and testing smaller physically motivated ML models,[147,148] models based on ridge regression (i.e., KRR), and multi-fidelity methods (which require a relatively large number of lower-quality data and a smaller number of higher-quality benchmark data for training). While NENCI-2021 on its own may be too small for training truly deep learning models (which require orders of magnitude more data, e.g., as one would find in the ANI-1[149] and QM7-X databases[150] of non-equilibrium conformations of small organic molecules), we also view this work as well as the composite databases (e.g., GMTKN55,[70] ACCDB,[76] and NCIAtlas[77–79]) and the very recently published datasets of Donchev *et al.*[80] as important steps in this direction.

We end this article with a brief discussion of several future research directions that could build off this work and potentially have an immediate impact in the field. For one, Paper II[91] in this series (in preparation) will critically assess the accuracy and reliability of a large number of popular DFT and WFT methods when describing the diverse array of non-equilibrium non-covalent interactions in NENCI-2021, thereby identifying the strengths and weaknesses of established first-principles methods. A simple and straightforward extension of NENCI-2021 would target dimers with increased intermolecular distances (e.g., beyond 1.1× the equilibrium separation), as benchmark E_{int} values for such complexes could play an important role in testing and training ML methods for predicting molecular multipoles[148] and polarizabilities,[151,152] as well as addressing important unresolved questions regarding the treatment of long-range electrostatics in ML-based potentials.[153] Another extension of NENCI-2021 could focus on simultaneously sampling non-equilibrium values for *both* the intra- and inter-molecular degrees of freedom in each of these molecular dimers using AIMD (or meta-dynamics to enhance sampling away from the global minimum).[100] Other important research thrusts would focus on expanding NENCI-2021 to further address the three challenges introduced above: (i) the need to describe NCIs in large molecular and condensed-phase systems can be addressed with extensions that focus on large/complex systems, higher-order molecular clusters (i.e., trimers, tetramers, etc.) to benchmark many-body effects/interactions,[71] and microsolvated complexes that explicitly include a variable number of solvent molecules; (ii) the need to describe the diverse types of NCIs on the same footing can be addressed by including NCI binding motifs that are either not found or underrepresented in NENCI-2021 (e.g., triple hydrogen bonds, quadrupole–quadrupole interactions, and ionic bonds); (iii) the need to describe NCIs in equilibrium and non-equilibrium systems on the same footing can be addressed by including complexes at more extreme (reduced and increased) intermolecular separations and angles and complexes between monomers in non-equilibrium configurations.

SUPPLEMENTARY MATERIAL

See the supplementary material for a single .zip file containing the Cartesian coordinates of the 7763 intermolecular complexes in NENCI-2021 (in .xyz format), a .csv file containing all of the CCSD(T)/CBS and SAPT energetic components (in kcal/mol), supplementary information for defining intermolecular interaction vectors, and an error analysis of alternative CCSD(T)/CBS extrapolation schemes.

ACKNOWLEDGMENTS

All the authors thank David Sherrill for helpful scientific discussions and Destiny Malloy for creating an early prototype of Fig. 1. All the authors acknowledge partial financial support from Cornell University through start-up funding. This work was supported by the National Science Foundation under Grant No. CHE-1945676. R.A.D. also acknowledges financial support from an Alfred P. Sloan Research Fellowship. This research used resources of the National Energy Research Scientific Computing Center, which is supported by the Office of Science of the U.S. Department of Energy under Contract No. DE-AC02-05CH11231.

AUTHOR DECLARATIONS

Conflicts of Interest

The authors have no conflicts to disclose.

Author Contributions

Z.M.S. and B.G.E. contributed equally to this work.

DATA AVAILABILITY

The data that support the findings of this study are available within the article and its supplementary material.

REFERENCES

[1] D. Langbein, *Theory of van der Waals Attraction* (Springer, Berlin, 1974).

[2] V. Parsegian, *Van der Waals Forces: A Handbook for Biologists, Chemists, Engineers, and Physicists* (Cambridge University Press, Cambridge, 2005).

[3] I. G. Kaplan, *Intermolecular Interactions: Physical Picture, Computational Methods and Model Potentials* (Wiley, New York, 2006).

[4] J. N. Israelachvili, *Intermolecular and Surface Forces* (Academic Press, Burlington, 2011).

[5] A. J. Stone, *The Theory of Intermolecular Forces*, 2nd ed. (Oxford University Press, Oxford, 2013).

[6] K. U. Wendt, K. Poralla, and G. E. Schulz, "Structure and function of a squalene cyclase," Science **277**, 1811–1815 (1997).

[7] Y. Zhao, Y. Domoto, E. Orentas, C. Beuchat, D. Emery, J. Mareda, N. Sakai, and S. Matile, "Catalysis with anion–π interactions," Angew. Chem., Int. Ed. **125**, 10124–10127 (2013).

[8] Y. Zhao, C. Beuchat, Y. Domoto, J. Gajewy, A. Wilson, J. Mareda, N. Sakai, and S. Matile, "Anion–π catalysis," J. Am. Chem. Soc. **136**, 2101–2111 (2014).

[9] C. R. Kennedy, S. Lin, and E. N. Jacobsen, "The cation–π interaction in small-molecule catalysis," Angew. Chem., Int. Ed. **55**, 12596–12624 (2016).

[10] Y. Zhao, Y. Cotelle, L. Liu, J. López-Andarias, A.-B. Bornhof, M. Akamatsu, N. Sakai, and S. Matile, "The emergence of anion–π catalysis," Acc. Chem. Res. **51**, 2255–2263 (2018).

[11] S. Yamada, "Cation–π interactions in organic synthesis," Chem. Rev. **118**, 11353–11432 (2018).

[12] P. Metrangolo, G. Resnati, T. Pilati, R. Liantonio, and F. Meyer, "Engineering functional materials by halogen bonding," J. Polym. Sci., Part A: Polym. Chem. **45**, 1–15 (2007).

[13] P. Metrangolo, T. Pilati, G. Terraneo, S. Biella, and G. Resnati, "Anion coordination and anion-templated assembly under halogen bonding control," CrystEngComm **11**, 1187–1196 (2009).

[14] C. W. Chen and H. W. Whitlock, Jr., "Molecular tweezers: A simple model of bifunctional intercalation," J. Am. Chem. Soc. **100**, 4921–4922 (1978).

[15] R. A. Bissell, E. Córdova, A. E. Kaifer, and J. F. Stoddart, "A chemically and electrochemically switchable molecular shuttle," Nature **369**, 133–137 (1994).

[16] V. Balzani, M. Gómez-López, and J. F. Stoddart, "Molecular machines," Acc. Chem. Res. **31**, 405–414 (1998).

[17] V. Balzani, A. Credi, F. M. Raymo, and J. F. Stoddart, "Artificial molecular machines," Angew. Chem., Int. Ed. **39**, 3348–3391 (2000).

[18] P. Kolb, D. M. Rosenbaum, J. J. Irwin, J. J. Fung, B. K. Kobilka, and B. K. Shoichet, "Structure-based discovery of β_2-adrenergic receptor ligands," Proc. Natl. Acad. Sci. U. S. A. **106**, 6843–6848 (2009).

[19] J. Fanfrlík, A. K. Bronowska, J. Řezáč, O. Přenosil, J. Konvalinka, and P. Hobza, "A reliable docking/scoring scheme based on the semiempirical quantum mechanical PM6-DH2 method accurately covering dispersion and H-bonding: HIV-1 protease with 22 ligands," J. Phys. Chem. B **114**, 12666–12678 (2010).

[20] C. Bissantz, B. Kuhn, and M. Stahl, "A medicinal chemist's guide to molecular interactions," J. Med. Chem. **53**, 5061–5084 (2010).

[21] B. Vorlová, D. Nachtigallová, J. Jirásková-Vaníčková, H. Ajani, P. Jansa, J. Řezáč, J. Fanfrlík, M. Otyepka, P. Hobza, J. Konvalinka *et al.*, "Malonate-based inhibitors of mammalian serine racemase: Kinetic characterization and structure-based computational study," Eur. J. Med. Chem. **89**, 189–197 (2015).

[22] K. A. Dill and J. L. MacCallum, "The protein-folding problem, 50 years on," Science **338**, 1042–1046 (2012).

[23] J. E. Jones, "On the determination of molecular fields. II. From the equation of state of a gas," Proc. R. Soc. London, Ser. A **106**, 463–477 (1924).

[24] J. W. Ponder, C. Wu, P. Ren, V. S. Pande, J. D. Chodera, M. J. Schnieders, I. Haque, D. L. Mobley, D. S. Lambrecht, R. A. DiStasio, Jr. *et al.*, "Current status of the AMOEBA polarizable force field," J. Phys. Chem. B **114**, 2549–2564 (2010).

[25] J. A. Lemkul, J. Huang, B. Roux, and A. D. MacKerell, Jr., "An empirical polarizable force field based on the classical Drude oscillator model: Development history and recent applications," Chem. Rev. **116**, 4983–5013 (2016).

[26] S. Grimme, "Density functional theory with London dispersion corrections," Wiley Interdiscip. Rev.: Comput. Mol. Sci. **1**, 211–228 (2011).

[27] K. Berland, V. R. Cooper, K. Lee, E. Schröder, T. Thonhauser, P. Hyldgaard, and B. I. Lundqvist, "van der Waals forces in density functional theory: A review of the vdW-DF method," Rep. Prog. Phys. **78**, 066501 (2015).

[28] J. Hermann, R. A. DiStasio, Jr., and A. Tkatchenko, "First-principles models for van der Waals interactions in molecules and materials: Concepts, theory, and applications," Chem. Rev. **117**, 4714–4758 (2017).

[29] C. Riplinger and F. Neese, "An efficient and near linear scaling pair natural orbital based local coupled cluster method," J. Chem. Phys. **138**, 034106 (2013).

[30] C. Riplinger, B. Sandhoefer, A. Hansen, and F. Neese, "Natural triple excitations in local coupled cluster calculations with pair natural orbitals," J. Chem. Phys. **139**, 134101 (2013).

[31] M. Rupp, A. Tkatchenko, K.-R. Müller, and O. A. von Lilienfeld, "Fast and accurate modeling of molecular atomization energies with machine learning," Phys. Rev. Lett. **108**, 058301 (2012).

[32] F. Brockherde, L. Vogt, L. Li, M. E. Tuckerman, K. Burke, and K.-R. Müller, "Bypassing the Kohn-Sham equations with machine learning," Nat. Commun. **8**, 872 (2017).

[33] K. T. Schütt, H. E. Sauceda, P.-J. Kindermans, A. Tkatchenko, and K.-R. Müller, "SchNet—A deep learning architecture for molecules and materials," J. Chem. Phys. **148**, 241722 (2018).

[34] T. Bereau, R. A. DiStasio, Jr., A. Tkatchenko, and O. A. von Lilienfeld, "Non-covalent interactions across organic and biological subsets of chemical space: Physics-based potentials parametrized from machine learning," J. Chem. Phys. **148**, 241706 (2018).

[35] D. P. Metcalf, A. Koutsoukas, S. A. Spronk, B. L. Claus, D. A. Loughney, S. R. Johnson, D. L. Cheney, and C. D. Sherrill, "Approaches for machine learning intermolecular interaction energies and application to energy components from symmetry adapted perturbation theory," J. Chem. Phys. **152**, 074103 (2020).

[36] A. M. Reilly, R. I. Cooper, C. S. Adjiman, S. Bhattacharya, A. D. Boese, J. G. Brandenburg, P. J. Bygrave, R. Bylsma, J. E. Campbell, R. Car *et al.*, "Report on the sixth blind test of organic crystal structure prediction methods," Acta Crystallogr., Sect. B: Struct. Sci., Cryst. Eng. Mater. **72**, 439–459 (2016).

[37] J. Hoja, H.-Y. Ko, M. A. Neumann, R. Car, R. A. DiStasio, Jr., and A. Tkatchenko, "Reliable and practical computational description of molecular crystal polymorphs," Sci. Adv. **5**, eaau3338 (2019).

[38] B. Cheng, G. Mazzola, C. J. Pickard, and M. Ceriotti, "Evidence for supercritical behaviour of high-pressure liquid hydrogen," Nature **585**, 217–220 (2020).

[39] S. Grimme, "Supramolecular binding thermodynamics by dispersion-corrected density functional theory," Chem. Eur. J. **18**, 9955–9964 (2012).

[40] A. Tkatchenko, R. A. DiStasio, Jr., R. Car, and M. Scheffler, "Accurate and efficient method for many-body van der Waals interactions," Phys. Rev. Lett. **108**, 236402 (2012).

[41] R. A. DiStasio, Jr., O. A. von Lilienfeld, and A. Tkatchenko, "Collective many-body van der Waals interactions in molecular systems," Proc. Natl. Acad. Sci. U. S. A. **109**, 14791–14795 (2012).

[42] R. Sedlak, T. Janowski, M. Pitoňák, J. Řezáč, P. Pulay, and P. Hobza, "Accuracy of quantum chemical methods for large noncovalent complexes," J. Chem. Theory Comput. **9**, 3364–3374 (2013).

[43] M. Goldey, R. A. DiStasio, Jr., Y. Shao, and M. Head-Gordon, "Shared memory multiprocessing implementation of resolution-of-the-identity second-order Møller–Plesset perturbation theory with attenuated and unattenuated results for intermolecular interactions between large molecules," Mol. Phys. **112**, 836–843 (2014).

[44] A. Ambrosetti, D. Alfè, R. A. DiStasio, Jr., and A. Tkatchenko, "Hard numbers for large molecules: Toward exact energetics for supramolecular systems," J. Phys. Chem. Lett. **5**, 849–855 (2014).

[45] R. A. DiStasio, Jr. and M. Head-Gordon, "Optimized spin-component scaled second-order Møller–Plesset perturbation theory for intermolecular interaction energies," Mol. Phys. **105**, 1073–1083 (2007).

[46] K. E. Riley and P. Hobza, "Assessment of the MP2 method, along with several basis sets, for the computation of interaction energies of biologically relevant hydrogen bonded and dispersion bound complexes," J. Phys. Chem. A **111**, 8257–8263 (2007).

[47] K. E. Riley, J. A. Platts, J. Řezáč, P. Hobza, and J. G. Hill, "Assessment of the performance of MP2 and MP2 variants for the treatment of noncovalent interactions," J. Phys. Chem. A **116**, 4159–4169 (2012).

[48] B. G. Ernst, K. U. Lao, A. G. Sullivan, and R. A. DiStasio, Jr., "Attracting opposites: Promiscuous ion–π binding in the nucleobases," J. Phys. Chem. A **124**, 4128–4140 (2020).

[49] Q. Wang, J. A. Rackers, C. He, R. Qi, C. Narth, L. Lagardere, N. Gresh, J. W. Ponder, J.-P. Piquemal, and P. Ren, "General model for treating short-range electrostatic penetration in a molecular mechanics force field," J. Chem. Theory Comput. **11**, 2609–2618 (2015).

[50] T. Gould, E. R. Johnson, and S. A. Tawfik, "Are dispersion corrections accurate outside equilibrium? A case study on benzene," Beilstein J. Org. Chem. **14**, 1181–1191 (2018).

[51] R. A. Mata and M. A. Suhm, "Benchmarking quantum chemical methods: Are we heading in the right direction?," Angew. Chem., Int. Ed. **56**, 11011–11018 (2017).

[52] J. Řezáč and P. Hobza, "Describing noncovalent interactions beyond the common approximations: How accurate is the 'gold standard,' CCSD(T) at the complete basis set limit?," J. Chem. Theory Comput. **9**, 2151–2155 (2013).

[53] L. A. Burns, M. S. Marshall, and C. D. Sherrill, "Appointing silver and bronze standards for noncovalent interactions: A comparison of spin-component-scaled (SCS), explicitly correlated (F12), and specialized wavefunction approaches," J. Chem. Phys. **141**, 234111 (2014).

[54] P. Jurečka, J. Šponer, J. Černý, and P. Hobza, "Benchmark database of accurate (MP2 and CCSD(T) complete basis set limit) interaction energies of small model complexes, DNA base pairs, and amino acid pairs," Phys. Chem. Chem. Phys. **8**, 1985–1993 (2006).

[55] M. S. Marshall, L. A. Burns, and C. D. Sherrill, "Basis set convergence of the coupled-cluster correction, $\delta^{\mathrm{MP2}}_{\mathrm{CCSD(T)}}$: Best practices for benchmarking non-covalent interactions and the attendant revision of the S22, NBC10, HBC6, and HSG databases," J. Chem. Phys. **135**, 194102 (2011).

[56] L. Gráfová, M. Pitoňák, J. Řezáč, and P. Hobza, "Comparative study of selected wave function and density functional methods for noncovalent interaction energy calculations using the extended S22 data set," J. Chem. Theory Comput. **6**, 2365–2376 (2010).

[57] J. Řezáč, K. E. Riley, and P. Hobza, "S66: A well-balanced database of benchmark interaction energies relevant to biomolecular structures," J. Chem. Theory Comput. **7**, 2427–2438 (2011).

[58] B. Brauer, M. K. Kesharwani, S. Kozuch, and J. M. L. Martin, "The S66x8 benchmark for noncovalent interactions revisited: Explicitly correlated *ab initio* methods and density functional theory," Phys. Chem. Chem. Phys. **18**, 20905–20925 (2016).

[59] J. Řezáč, K. E. Riley, and P. Hobza, "Extensions of the S66 data set: More accurate interaction energies and angular-displaced nonequilibrium geometries," J. Chem. Theory Comput. **7**, 3466–3470 (2011).

[60] J. Řezáč, K. E. Riley, and P. Hobza, "Benchmark calculations of noncovalent interactions of halogenated molecules," J. Chem. Theory Comput. **8**, 4285–4292 (2012).

[61] T. M. Parker and C. D. Sherrill, "Assessment of empirical models versus high-accuracy ab initio methods for nucleobase stacking: Evaluating the importance of charge penetration," J. Chem. Theory Comput. **11**, 4197–4204 (2015).

[62] Y. Zhao and D. G. Truhlar, "Benchmark databases for nonbonded interactions and their use to test density functional theory," J. Chem. Theory Comput. **1**, 415–432 (2005).

[63] S. Tsuzuki, K. Honda, T. Uchimaru, and M. Mikami, "Estimated MP2 and CCSD(T) interaction energies of *n*-alkane dimers at the basis set limit: Comparison of the methods of Helgaker *et al.* and Feller," J. Chem. Phys. **124**, 114304 (2006).

[64] S. Kozuch and J. M. L. Martin, "Halogen bonds: Benchmarks and theoretical analysis," J. Chem. Theory Comput. **9**, 1918–1931 (2013).

[65] B. J. Mintz and J. M. Parks, "Benchmark interaction energies for biologically relevant noncovalent complexes containing divalent sulfur," J. Phys. Chem. A **116**, 1086–1092 (2012).

[66] J. Řezáč and P. Hobza, "Advanced corrections of hydrogen bonding and dispersion for semiempirical quantum mechanical methods," J. Chem. Theory Comput. **8**, 141–151 (2012).

[67] J. C. Faver, M. L. Benson, X. He, B. P. Roberts, B. Wang, M. S. Marshall, M. R. Kennedy, C. D. Sherrill, and K. M. Merz, "Formal estimation of errors in computed absolute interaction energies of protein-ligand complexes," J. Chem. Theory Comput. **7**, 790–797 (2011).

[68] J. Řezáč, M. Dubecký, P. Jurečka, and P. Hobza, "Extensions and applications of the A24 data set of accurate interaction energies," Phys. Chem. Chem. Phys. **17**, 19268–19277 (2015).

[69] D. G. A. Smith, L. A. Burns, K. Patkowski, and C. D. Sherrill, "Revised damping parameters for the D3 dispersion correction to density functional theory," J. Phys. Chem. Lett. **7**, 2197–2203 (2016).

[70] L. Goerigk, A. Hansen, C. Bauer, S. Ehrlich, A. Najibi, and S. Grimme, "A look at the density functional theory zoo with the advanced GMTKN55 database for general main group thermochemistry, kinetics and noncovalent interactions," Phys. Chem. Chem. Phys. **19**, 32184–32215 (2017).

[71] J. Řezáč, Y. Huang, P. Hobza, and G. J. O. Beran, "Benchmark calculations of three-body intermolecular interactions and the performance of low-cost electronic structure methods," J. Chem. Theory Comput. **11**, 3065–3079 (2015).

[72] J. G. McDaniel and J. R. Schmidt, "Physically-motivated force fields from symmetry-adapted perturbation theory," J. Phys. Chem. A **117**, 2053–2066 (2013).

[73] J. G. McDaniel and J. R. Schmidt, "First-principles many-body force fields from the gas phase to liquid: A 'universal' approach," J. Phys. Chem. B **118**, 8042–8053 (2014).

[74] S. Vandenbrande, M. Waroquier, V. V. Speybroeck, and T. Verstraelen, "The monomer electron density force field (MEDFF): A physically inspired model for noncovalent interactions," J. Chem. Theory Comput. **13**, 161–179 (2017).

[75] J. Řezáč and P. Hobza, "Benchmark calculations of interaction energies in noncovalent complexes and their applications," Chem. Rev. **116**, 5038–5071 (2016).

[76] P. Morgante and R. Peverati, "ACCDB: A collection of chemistry databases for broad computational purposes," J. Comput. Chem. **40**, 839–848 (2019).

[77] J. Řezáč, "Non-covalent interactions atlas benchmark data sets: Hydrogen bonding," J. Chem. Theory Comput. **16**, 2355–2368 (2020).

[78] J. Řezáč, "Non-covalent interactions atlas benchmark data sets 2: Hydrogen bonding in an extended chemical space," J. Chem. Theory Comput. **16**, 6305–6316 (2020).

[79] K. Kříž, M. Nováček, and J. Řezáč, "Non-covalent interactions atlas benchmark data sets 3: Repulsive contacts," J. Chem. Theory Comput. **17**, 1548–1561 (2021).

[80] A. G. Donchev, A. G. Taube, E. Decolvenaere, C. Hargus, R. T. McGibbon, K.-H. Law, B. A. Gregersen, J.-L. Li, K. Palmo, K. Siva, M. Bergdorf, J. L. Klepeis, and D. E. Shaw, "Quantum chemical benchmark databases of gold-standard dimer interaction energies," Sci. Data **8**, 55 (2021).

[81] V. M. Miriyala and J. Řezáč, "Testing semiempirical quantum mechanical methods on a data set of interaction energies mapping repulsive contacts in organic molecules," J. Phys. Chem. A **122**, 2801–2808 (2018).

[82] V. Bazgier, K. Berka, M. Otyepka, and P. Banáš, "Exponential repulsion improves structural predictability of molecular docking," J. Comput. Chem. **37**, 2485–2494 (2016).

[83] L. Song, N. Fu, B. G. Ernst, W. H. Lee, M. O. Frederick, R. A. DiStasio, Jr., and S. Lin, "Dual electrocatalysis enables enantioselective hydrocyanation of conjugated alkenes," Nat. Chem. **12**, 747–754 (2020).

[84] C. J. Sahle, C. Sternemann, C. Schmidt, S. Lehtola, S. Jahn, L. Simonelli, S. Huotari, M. Hakala, T. Pylkkänen, A. Nyrow, K. Mende, M. Tolan, K. Hämäläinen, and M. Wilke, "Microscopic structure of water at elevated pressures and temperatures," Proc. Natl. Acad. Sci. U. S. A. **110**, 6301–6306 (2013).

[85] M. Miao, Y. Sun, E. Zurek, and H. Lin, "Chemistry under high pressure," Nat. Rev. Chem. **4**, 508–527 (2020).

[86] H. Liu, I. I. Naumov, R. Hoffmann, N. W. Ashcroft, and R. J. Hemley, "Potential high-T_c superconducting lanthanum and yttrium hydrides at high pressure," Proc. Natl. Acad. Sci. U. S. A. **114**, 6990–6995 (2017).

[87] E. G. Hohenstein, J. Duan, and C. D. Sherrill, "Origin of the surprising enhancement of electrostatic energies by electron-donating substituents in substituted sandwich benzene dimers," J. Am. Chem. Soc. **133**, 13244–13247 (2011).

[88] B. Jeziorski, R. Moszynski, and K. Szalewicz, "Perturbation theory approach to intermolecular potential energy surfaces of van der Waals complexes," Chem. Rev. **94**, 1887–1930 (1994).

[89] K. U. Lao and J. M. Herbert, "Breakdown of the single-exchange approximation in third-order symmetry-adapted perturbation theory," J. Phys. Chem. A **116**, 3042–3047 (2012).

[90] N. Mardirossian and M. Head-Gordon, "Thirty years of density functional theory in computational chemistry: An overview and extensive assessment of 200 density functionals," Mol. Phys. **115**, 2315–2372 (2017).

[91] B. G. Ernst, Z. M. Sparrow, and R. A. DiStasio, Jr., "NENCI-2021 Part II: Evaluating the performance of quantum chemical approximations on the NENCI-2021 benchmark database" (unpublished).

[92] P. Politzer, J. S. Murray, and T. Clark, "Halogen bonding: An electrostatically-driven highly directional noncovalent interaction," Phys. Chem. Chem. Phys. **12**, 7748–7757 (2010).

[93] D. A. Dougherty, "Cation–π interactions in chemistry and biology: A new view of benzene, Phe, Tyr, and Trp," Science **271**, 163–168 (1996).

[94] J. C. Ma and D. A. Dougherty, "The cation–π interaction," Chem. Rev. **97**, 1303–1324 (1997).

[95] A. S. Mahadevi and G. N. Sastry, "Cation–π interaction: Its role and relevance in chemistry, biology, and material science," Chem. Rev. **113**, 2100–2138 (2012).

[96] A. Frontera, P. Gamez, M. Mascal, T. J. Mooibroek, and J. Reedijk, "Putting anion–π interactions into perspective," Angew. Chem., Int. Ed. **50**, 9564–9583 (2011).

[97] B. L. Schottel, H. T. Chifotides, and K. R. Dunbar, "Anion–π interactions," Chem. Soc. Rev. **37**, 68–83 (2008).

[98] M. S. Marshall, R. P. Steele, K. S. Thanthiriwatte, and C. D. Sherrill, "Potential energy curves for cation–π interactions: Off-axis configurations are also attractive," J. Phys. Chem. A **113**, 13628–13632 (2009).

[99] J. Novotnỳ, S. Bazzi, R. Marek, and J. Kozelka, "Lone-pair–π interactions: Analysis of the physical origin and biological implications," Phys. Chem. Chem. Phys. **18**, 19472–19481 (2016).

[100] A. Laio and M. Parrinello, "Escaping free-energy minima," Proc. Natl. Acad. Sci. U. S. A. **99**, 12562–12566 (2002).

[101] S. F. Boys and F. Bernardi, "The calculation of small molecular interactions by the differences of separate total energies. Some procedures with reduced errors," Mol. Phys. **19**, 553–566 (1970).

[102] T. H. Dunning, Jr., "Gaussian basis sets for use in correlated molecular calculations. I. The atoms boron through neon and hydrogen," J. Chem. Phys. **90**, 1007–1023 (1989).

[103] R. A. Kendall, T. H. Dunning, Jr., and R. J. Harrison, "Electron affinities of the first-row atoms revisited. Systematic basis sets and wave functions," J. Chem. Phys. **96**, 6796–6806 (1992).

[104] D. E. Woon and T. H. Dunning, Jr., "Gaussian basis sets for use in correlated molecular calculations. III. The atoms aluminum through argon," J. Chem. Phys. **98**, 1358–1371 (1993).

[105] A. K. Wilson, D. E. Woon, K. A. Peterson, and T. H. Dunning, Jr., "Gaussian basis sets for use in correlated molecular calculations. IX. The atoms gallium through krypton," J. Chem. Phys. **110**, 7667–7676 (1999).

[106] B. P. Prascher, D. E. Woon, K. A. Peterson, T. H. Dunning, Jr., and A. K. Wilson, "Gaussian basis sets for use in correlated molecular calculations. VII. Valence, core-valence, and scalar relativistic basis sets for Li, Be, Na, and Mg," Theor. Chem. Acc. **128**, 69–82 (2011).

[107] P. Jurečka, P. Nachtigall, and P. Hobza, "RI-MP2 calculations with extended basis sets—A promising tool for study of H-bonded and stacked DNA base pairs," Phys. Chem. Chem. Phys. **3**, 4578–4582 (2001).

[108] A. E. DePrince and C. D. Sherrill, "Accuracy and efficiency of coupled-cluster theory using density fitting/Cholesky decomposition, frozen natural orbitals, and a t_1-transformed Hamiltonian," J. Chem. Theory Comput. **9**, 2687–2696 (2013).

[109] F. Weigend, "A fully direct RI-HF algorithm: Implementation, optimised auxiliary basis sets, demonstration of accuracy and efficiency," Phys. Chem. Chem. Phys. **4**, 4285–4291 (2002).

[110] F. Weigend, A. Köhn, and C. Hättig, "Efficient use of the correlation consistent basis sets in resolution of the identity MP2 calculations," J. Chem. Phys. **116**, 3175–3183 (2002).

[111] C. Hättig, "Optimization of auxiliary basis sets for RI-MP2 and RI-CC2 calculations: Core–valence and quintuple-ζ basis sets for H to Ar and QZVPP basis sets for Li to Kr," Phys. Chem. Chem. Phys. **7**, 59–66 (2005).

[112] F. Weigend, "Hartree–Fock exchange fitting basis sets for H to Rn," J. Comput. Chem. **29**, 167–175 (2008).

[113] A. Hellweg and D. Rappoport, "Development of new auxiliary basis functions of the Karlsruhe segmented contracted basis sets including diffuse basis functions (def2-SVPD, def2-TZVPPD, and def2-QVPPD) for RI-MP2 and RI-CC calculations," Phys. Chem. Chem. Phys. **17**, 1010–1017 (2015).

[114] H.-J. Werner, P. J. Knowles, G. Knizia, F. R. Manby, and M. Schütz, "Molpro: A general-purpose quantum chemistry program package," Wiley Interdiscip. Rev.: Comput. Mol. Sci. **2**, 242–253 (2012).

[115] H.-J. Werner, P. J. Knowles, G. Knizia, F. R. Manby, M. Schütz, P. Celani, W. Györffy, D. Kats, T. Korona, R. Lindh, A. Mitrushenkov, G. Rauhut, K. R. Shamasundar, T. B. Adler, R. D. Amos, S. J. Bennie, A. Bernhardsson, A. Berning, D. L. Cooper, M. J. O. Deegan, A. J. Dobbyn, F. Eckert, E. Goll, C. Hampel, A. Hesselmann, G. Hetzer, T. Hrenar, G. Jansen, C. Köppl, S. J. R. Lee, Y. Liu, A. W. Lloyd,

Q. Ma, R. A. Mata, A. J. May, S. J. McNicholas, W. Meyer, T. F. Miller III, M. E. Mura, A. Nicklass, D. P. O'Neill, P. Palmieri, D. Peng, K. Pflüger, R. Pitzer, M. Reiher, T. Shiozaki, H. Stoll, A. J. Stone, R. Tarroni, T. Thorsteinsson, M. Wang, and M. Welborn, Molpro, version 2019.2, a package of ab initio programs, 2019, see https://www.molpro.net.

[116] E. Papajak, J. Zheng, X. Xu, H. R. Leverentz, and D. G. Truhlar, "Perspectives on basis sets beautiful: Seasonal plantings of diffuse basis functions," J. Chem. Theory Comput. **7**, 3027–3034 (2011).

[117] A. Halkier, T. Helgaker, P. Jørgensen, W. Klopper, H. Koch, J. Olsen, and A. K. Wilson, "Basis-set convergence in correlated calculations on Ne, N_2, and H_2O," Chem. Phys. Lett. **286**, 243–252 (1998).

[118] E. G. Hohenstein and C. D. Sherrill, "Density fitting and Cholesky decomposition approximations in symmetry-adapted perturbation theory: Implementation and application to probe the nature of π–π interactions in linear acenes," J. Chem. Phys. **132**, 184111 (2010).

[119] E. G. Hohenstein, R. M. Parrish, C. D. Sherrill, J. M. Turney, and H. F. Schaefer III, "Large-scale symmetry-adapted perturbation theory computations via density fitting and Laplace transformation techniques: Investigating the fundamental forces of DNA-intercalator interactions," J. Chem. Phys. **135**, 174107 (2011).

[120] E. G. Hohenstein and C. D. Sherrill, "Density fitting of intramonomer correlation effects in symmetry-adapted perturbation theory," J. Chem. Phys. **133**, 014101 (2010).

[121] E. G. Hohenstein and C. D. Sherrill, "Efficient evaluation of triple excitations in symmetry-adapted perturbation theory via second-order Møller–Plesset perturbation theory natural orbitals," J. Chem. Phys. **133**, 104107 (2010).

[122] T. M. Parker, L. A. Burns, R. M. Parrish, A. G. Ryno, and C. D. Sherrill, "Levels of symmetry adapted perturbation theory (SAPT). I. Efficiency and performance for interaction energies," J. Chem. Phys. **140**, 094106 (2014).

[123] R. M. Parrish, L. A. Burns, D. G. A. Smith, A. C. Simmonett, A. E. DePrince, E. G. Hohenstein, U. Bozkaya, A. Y. Sokolov, R. Di Remigio, R. M. Richard, J. F. Gonthier, A. M. James, H. R. McAlexander, A. Kumar, M. Saitow, X. Wang, B. P. Pritchard, P. Verma, H. F. Schaefer III, K. Patkowski, R. A. King, E. F. Valeev, F. A. Evangelista, J. M. Turney, T. D. Crawford, and C. D. Sherrill, "Psi4 1.1: An open-source electronic structure program emphasizing automation, advanced libraries, and interoperability," J. Chem. Theory Comput. **13**, 3185–3197 (2017).

[124] A. Bondi, "van der Waals volumes and radii," J. Chem. Phys. **68**, 441–451 (1964).

[125] R. S. Rowland and R. Taylor, "Intermolecular nonbonded contact distances in organic crystal structures: Comparison with distances expected from van der Waals radii," J. Phys. Chem. **100**, 7384–7391 (1996).

[126] N. J. Singh, S. K. Min, D. Y. Kim, and K. S. Kim, "Comprehensive energy analysis for various types of π-interaction," J. Chem. Theory Comput. **5**, 515–529 (2009).

[127] K. S. Thanthiriwatte, E. G. Hohenstein, L. A. Burns, and C. D. Sherrill, "Assessment of the performance of DFT and DFT-D methods for describing distance dependence of hydrogen-bonded interactions," J. Chem. Theory Comput. **7**, 88–96 (2011).

[128] M. O. Sinnokrot and C. D. Sherrill, "High-accuracy quantum mechanical studies of π–π interactions in benzene dimers," J. Phys. Chem. A **110**, 10656–10668 (2006).

[129] A. L. Ringer, M. S. Figgs, M. O. Sinnokrot, and C. D. Sherrill, "Aliphatic C–H/π interactions: Methane–benzene, methane–phenol, and methane–indole complexes," J. Phys. Chem. A **110**, 10822–10828 (2006).

[130] C. D. Sherrill, T. Takatani, and E. G. Hohenstein, "An assessment of theoretical methods for nonbonded interactions: Comparison to complete basis set limit coupled-cluster potential energy curves for the benzene dimer, the methane dimer, benzene-methane, and benzene-H_2S," J. Phys. Chem. A **113**, 10146–10159 (2009).

[131] Y. Geng, T. Takatani, E. G. Hohenstein, and C. D. Sherrill, "Accurately characterizing the π–π interaction energies of indole-benzene complexes," J. Phys. Chem. A **114**, 3576–3582 (2010).

[132] K. U. Lao, R. Schäffer, G. Jansen, and J. M. Herbert, "Accurate description of intermolecular interactions involving ions using symmetry-adapted perturbation theory," J. Chem. Theory Comput. **11**, 2473–2486 (2015).

[133] L. A. Burns, J. C. Faver, Z. Zheng, M. S. Marshall, D. G. A. Smith, K. Vanommeslaeghe, A. D. MacKerell, K. M. Merz, and C. D. Sherrill, "The biofragment database (BFDb): An open-data platform for computational chemistry analysis of noncovalent interactions," J. Chem. Phys. **147**, 161727 (2017).

[134] J. A. Rackers, Q. Wang, C. Liu, J.-P. Piquemal, P. Ren, and J. W. Ponder, "An optimized charge penetration model for use with the AMOEBA force field," Phys. Chem. Chem. Phys. **19**, 276–291 (2017).

[135] A. Halkier, T. Helgaker, P. Jørgensen, W. Klopper, and J. Olsen, "Basis-set convergence of the energy in molecular Hartree–Fock calculations," Chem. Phys. Lett. **302**, 437–446 (1999).

[136] A. L. L. East and W. D. Allen, "The heat of formation of NCO," J. Chem. Phys. **99**, 4638–4650 (1993).

[137] M. O. Sinnokrot, E. F. Valeev, and C. D. Sherrill, "Estimates of the ab initio limit for π–π interactions: The benzene dimer," J. Am. Chem. Soc. **124**, 10887–10893 (2002).

[138] M. Pitoňák, T. Janowski, P. Neogrády, P. Pulay, and P. Hobza, "Convergence of the CCSD(T) correction term for the stacked complex methyl adenine-methyl thymine: Comparison with lower-cost alternatives," J. Chem. Theory Comput. **5**, 1761–1766 (2009).

[139] A. D. Boese, J. M. L. Martin, and W. Klopper, "Basis set limit coupled cluster study of H-bonded systems and assessment of more approximate methods," J. Phys. Chem. A **111**, 11122–11133 (2007).

[140] L. Demovičová, P. Hobza, and J. Řezáč, "Evaluation of composite schemes for CCSDT(Q) calculations of interaction energies of noncovalent complexes," Phys. Chem. Chem. Phys. **16**, 19115–19121 (2014).

[141] J. E. Deustua, J. Shen, and P. Piecuch, "Converging high-level coupled-cluster energetics by Monte Carlo sampling and moment expansions," Phys. Rev. Lett. **119**, 223003 (2017).

[142] C. J. C. Scott, R. Di Remigio, T. D. Crawford, and A. J. W. Thom, "Diagrammatic coupled cluster Monte Carlo," J. Phys. Chem. Lett. **10**, 925–935 (2019).

[143] C. J. C. Scott, R. Di Remigio, T. D. Crawford, and A. J. W. Thom, "Theory and implementation of a novel stochastic approach to coupled cluster," J. Chem. Phys. **153**, 144117 (2020).

[144] N. S. Blunt, S. D. Smart, J. A. F. Kersten, J. S. Spencer, G. H. Booth, and A. Alavi, "Semi-stochastic full configuration interaction quantum Monte Carlo: Developments and application," J. Chem. Phys. **142**, 184107 (2015).

[145] S. Sharma, A. A. Holmes, G. Jeanmairet, A. Alavi, and C. J. Umrigar, "Semistochastic heat-bath configuration interaction method: Selected configuration interaction with semistochastic perturbation theory," J. Chem. Theory Comput. **13**, 1595–1604 (2017).

[146] J. Li, M. Otten, A. A. Holmes, S. Sharma, and C. J. Umrigar, "Fast semistochastic heat-bath configuration interaction," J. Chem. Phys. **149**, 214110 (2018).

[147] J. Proppe, S. Gugler, and M. Reiher, "Gaussian process-based refinement of dispersion corrections," J. Chem. Theory Comput. **15**, 6046–6060 (2019).

[148] M. Veit, D. M. Wilkins, Y. Yang, R. A. DiStasio, Jr., and M. Ceriotti, "Predicting molecular dipole moments by combining atomic partial charges and atomic dipoles," J. Chem. Phys. **153**, 024113 (2020).

[149] J. S. Smith, O. Isayev, and A. E. Roitberg, "ANI-1, a data set of 20 million calculated off-equilibrium conformations for organic molecules," Sci. Data **4**, 170193 (2017).

[150] J. Hoja, L. M. Sandonas, B. G. Ernst, A. Vazquez-Mayagoitia, R. A. DiStasio, Jr., and A. Tkatchenko, "QM7-X, a comprehensive dataset of quantum-mechanical properties spanning the chemical space of small organic molecules," Sci. Data **8**, 43 (2021).

[151] Y. Yang, K. U. Lao, D. M. Wilkins, A. Grisafi, M. Ceriotti, and R. A. DiStasio, Jr., "Quantum mechanical static dipole polarizabilities in the QM7b and AlphaML showcase databases," Sci. Data **6**, 152 (2019).

[152] D. M. Wilkins, A. Grisafi, Y. Yang, K. U. Lao, R. A. DiStasio, Jr., and M. Ceriotti, "Accurate molecular polarizabilities with coupled cluster theory and machine learning," Proc. Natl. Acad. Sci. U. S. A. **116**, 3401–3406 (2019).

[153] S. Yue, M. C. Muniz, M. F. Calegari Andrade, L. Zhang, R. Car, and A. Z. Panagiotopoulos, "When do short-range atomistic machine-learning models fall short?," J. Chem. Phys. **154**, 034111 (2021).