Journal of Medicinal Chemistry

pubs.acs.org/jmc

Article

A Distributional Model of Bound Ligand Conformational Strain: From Small Molecules up to Large Peptidic Macrocycles

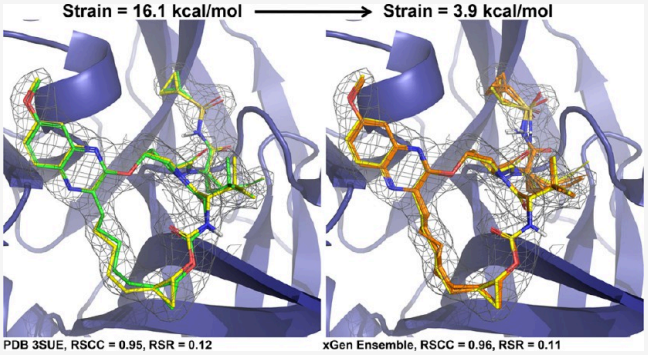
Ajay N. Jain,* Alexander C. Brueckner, Ann E. Cleves, Mikhail Reibarkh, and Edward C. Sherer*

Cite This: *J. Med. Chem.* 2023, 66, 1955−1971

Read Online

ACCESS | Metrics & More | Article Recommendations | Supporting Information

ABSTRACT: The internal conformational strain incurred by ligands upon binding a target site has a critical impact on binding affinity, and expectations about the magnitude of ligand strain guide conformational search protocols. Estimates for bound ligand strain begin with modeled ligand atomic coordinates from X-ray co-crystal structures. By deriving low-energy conformational ensembles to fit X-ray diffraction data, calculated strain energies are substantially reduced compared with prior approaches. We show that the distribution of expected global strain energy values is dependent on molecular size in a superlinear manner. The distribution of strain energy follows a rectified normal distribution whose mean and variance are related to conformational complexity. The modeled strain distribution closely matches calculated strain values from experimental data comprising over 3000 protein−ligand complexes. The distributional model has direct implications for conformational search protocols as well as for directions in molecular design.



INTRODUCTION

Estimating bound ligand strain is complicated by the fact that the tools available for X-ray crystallography model refinement are better developed for protein modeling than for ligand modeling. Very often, the modeled ligand coordinates yield very high energy values, whether calculated by force field or quantum mechanics (QM) approaches. This has been established in a number of studies and reviews concerned with estimating strain energy[1−10] and studies and perspectives involving X-ray model accuracy.[11−13] For the strain-focused studies, especially the earlier ones (e.g., the influential work of Perola and Charifson[3]), unrealistically high strain energies were frequently calculated, despite employing means to overcome the limitations of the modeled X-ray ligand coordinates. In perspectives on X-ray ligand modeling accuracy, frequent and often obviously incorrect ligand geometries have been well documented. Liebeschuetz et al.[11] found that a majority of PDB complex ligands showed evidence of incorrect bond lengths and angles being used in refinement, and a quarter of ligand structures had avoidable geometric errors that were potentially large enough to lead to a mischaracterization of binding interactions. Subsequent work highlighted continuing problems with ligand fit, even in PDB complexes of moderate or high resolution.[12]

Reynolds[13] summarized as follows: except for very high-resolution cases, structures are fitted models that cannot be assigned using the experimental density information alone, and refinement protocols often require that the crystallographer be responsible for determining appropriate structural constraints for the ligand, which can be challenging. Improvements in protocols for ligand refinement have been developed, including AFITT, qFit-Ligand, and DivCon,[14−19] but such methods have not been adopted widely enough to make an impact on the vast majority of publicly available protein−ligand structure data. Consequently, a universal aspect of strain estimation for bound ligands is the use of a "surrogate-energy conformer" in place of a conformer with crystallographically modeled atomic coordinates. Methods for deriving the surrogate-energy conformer and for evaluating its energy vary widely, though all seek to identify an energetically reasonable surrogate whose deviations from deposited ligand coordinates are minimal.

We recently introduced a real-space ligand refinement method (called "xGen") suitable for application to typical small molecules but also capable of efficient refinement for synthetic and peptidic macrocycles.[20,21] Two features of the method are critical in the context of ligand strain. First, the method explicitly seeks low-energy solutions to fitting X-ray density, making use of a variant of MMFF94s, and it produces energy-surrogate conformers as part of the fitting process. Second, rather than employing atom-specific B-factors to account for atomic positional uncertainty, occupancy-weighed

Received: October 27, 2022
Published: January 26, 2023

ACS Publications

© 2023 The Authors. Published by American Chemical Society

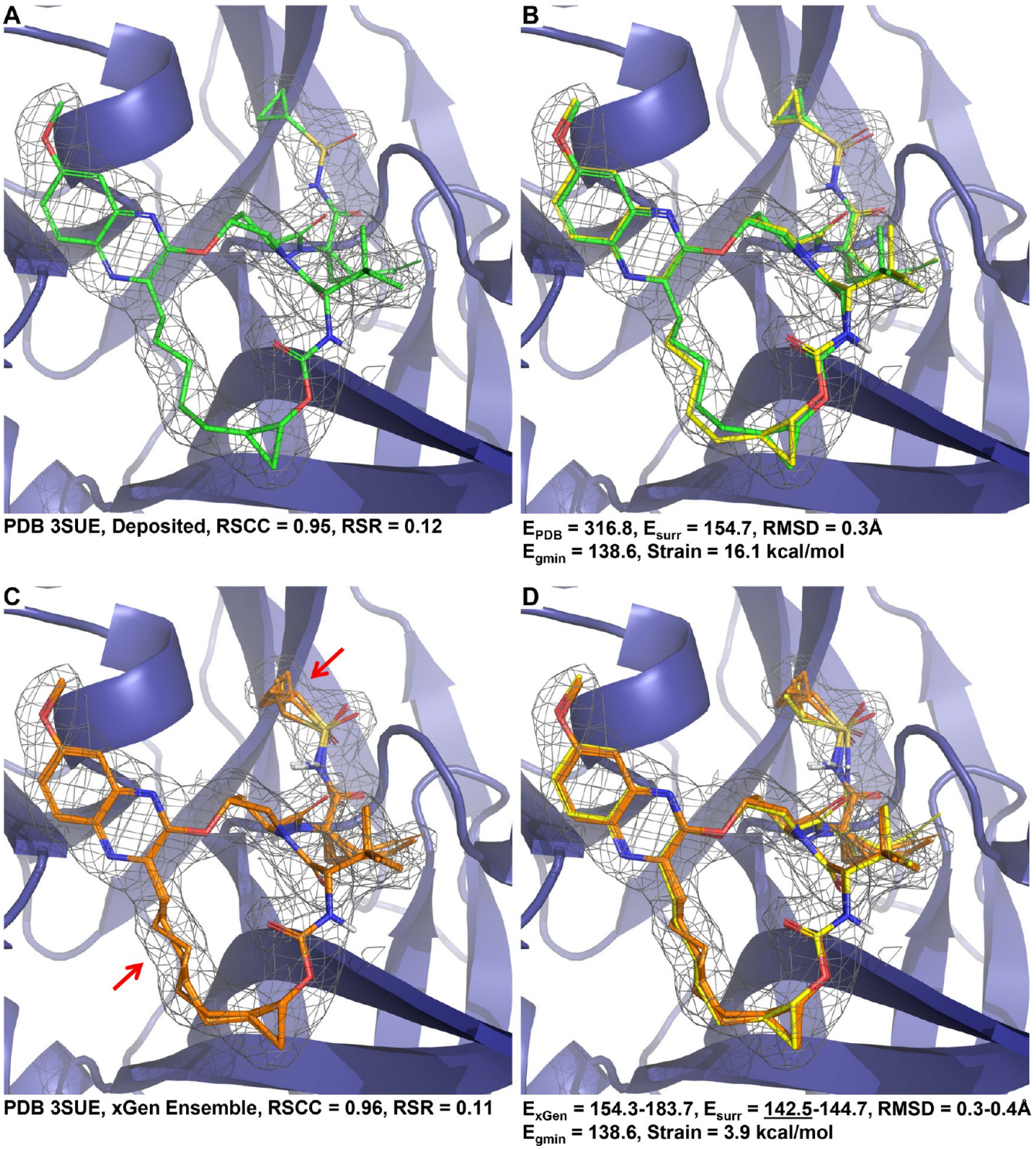


Figure 1. Alternative methods for estimating bound ligand strain for grazoprevir bound to NS3/4A protease variant R155K. (A) Deposited ligand model coordinates (green), showing the experimental electron density contour from the $2|F_o| - |F_c|$ map at 1.0σ, with real-space fit metrics calculated using the modeled atom-specific B-factors. (B) The energy-surrogate conformer (yellow), with force field energy values for the modeled coordinates, the energy-surrogate, the global minimum, and the calculated strain. (C) Conformer ensemble from xGen (orange), with real-space fit metrics calculated using a constant B-factor and conformer-level occupancy weighting. (D) The lowest-energy xGen surrogate conformer (yellow), with the range of energy values for the ensemble and energy-surrogates along with the calculated strain.

conformational ensembles are sought such that their joint contributions to calculated electron density fit the experimental density well. The method simultaneously yielded improvements in real-space electron density fit and significant reductions in nominal bound ligand energies.

Figure 1 shows the contrast between a standard approach to ligand modeling and strain estimation and the xGen-based approach. Panel A shows the deposited ligand coordinates for grazoprevir bound to an NS3/4A protease variant, which exhibited excellent fidelity to the experimental density by both

https://doi.org/10.1021/acs.jmedchem.2c01744
J. Med. Chem. 2023, 66, 1955−1971

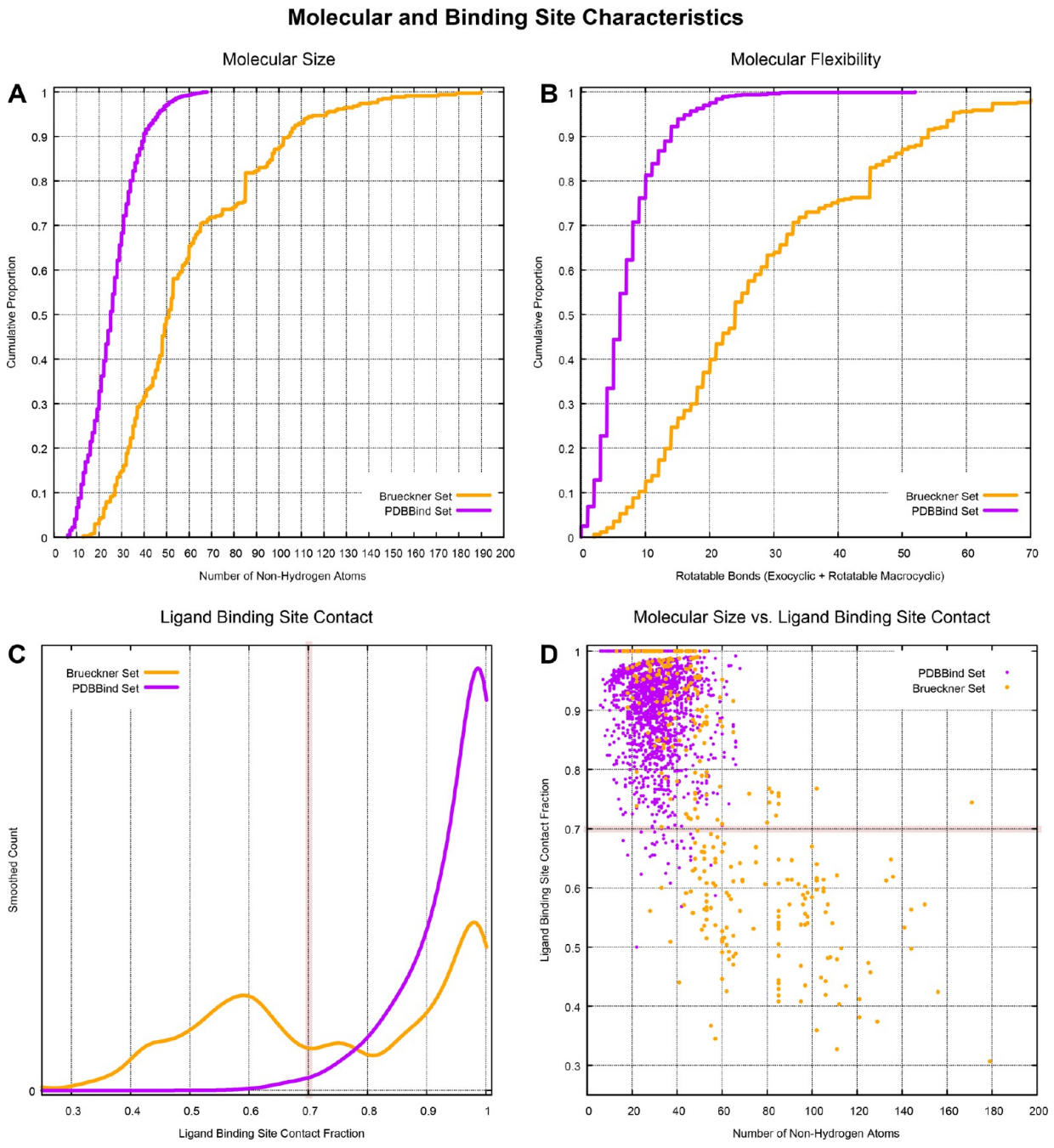


Figure 2. Molecular size, flexibility, and degree of ligand binding site contact. (A) Cumulative histogram of molecular size. (B) Cumulative histogram of total molecular flexibility, which includes exocyclic rotatable bonds plus rotatable macrocyclic bonds. (C) Smoothed histogram of the ligand binding contact, defined as the fraction of ligand atoms within 1.0 Å of any protein atom (distance measured between VdW surfaces). (D) Relationship between molecular size and binding site contact.

RSCC and RSR (real-space metrics for quality of density fit[22]). Panel B shows the energy-surrogate conformer (yellow) obtained by position-constrained minimization using a square-welled quadratic positional restraint on non-hydrogen atoms (1.0 kcal/mol/Å^2 for deviations beyond 0.1 Å). Positionally restrained minimization reduced the nominal energy of the deposited bound ligand model from 316.8 (labeled as E_{PDB}) to 154.7 kcal/mol (labeled as E_{surr}), with minor geometric deviation, illustrating the need for an energy-surrogate conformer in place of directly modeled ligand coordinates. Global strain is calculated based on the difference between the bound-state energy (E_{surr}) and the unbound-state minimum energy (E_{gmin}), which is the global minimum energy from an exhaustive conformational search of the ligand.

Panel C shows the four-conformer xGen ensemble, which differed from the deposited ligand in the macrocyclic linker and in the terminal cyclopropyl, allowing for some movement in both areas (marked by red arrows). The xGen ensemble is both a good fit to the density and low in energy. Panel D shows the ensemble conformers (orange) and energy-surrogate conformer (yellow) with the lowest energy. Energy ranges for the ensemble conformers and corresponding energy-surrogates are labeled as E_{xGen} and E_{surr}. Estimated strain energy was reduced from +16.1 to +3.9 kcal/mol, which seems more plausible in light of the

https://doi.org/10.1021/acs.jmedchem.2c01744
J. Med. Chem. 2023, 66, 1955−1971

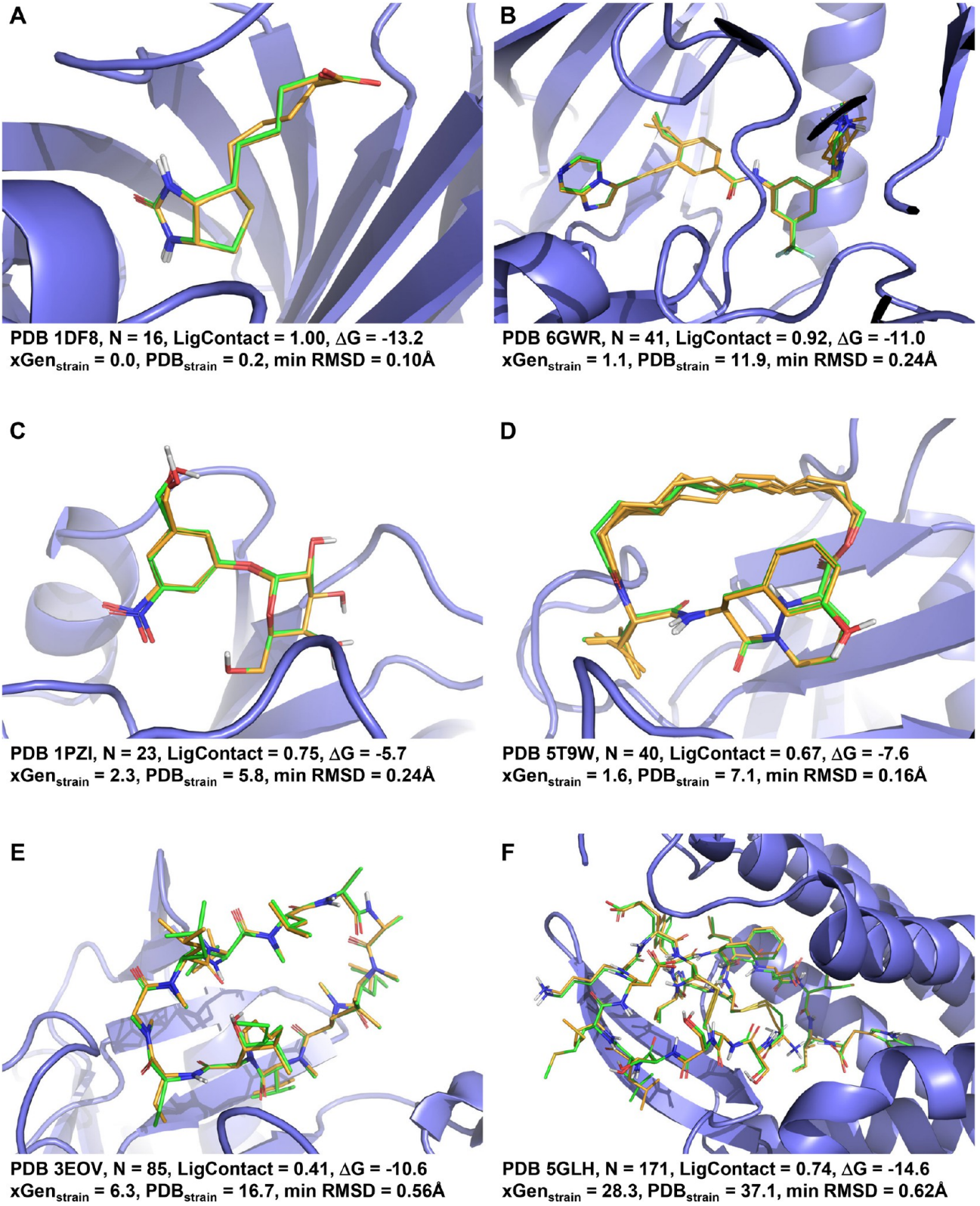


Figure 3. Examples spanning different molecular sizes and ligand binding site contact (energy values in kcal/mol). PDB ligand coordinates are shown with green carbons, and xGen fitted ensembles are shown with orange carbons. (A) Biotin bound to the S45A mutant of streptavidin. (B) Inhibitor bound to the kinase domain of human DDR1. (C) Antagonist bound to heat-labile Enterotoxin B. (D) Sanglifehrin A analog bound to cyclophilin. (E) Cyclosporin A bound to a *Leishmania donovani* cyclophilin (ΔG from Venugopal et al.[26]). (F) Endothelin in complex with human endothelin receptor type-B (ΔG from Shihoya et al.[27]).

https://doi.org/10.1021/acs.jmedchem.2c01744
J. Med. Chem. 2023, 66, 1955−1971

approximate free energy of binding being −12.3 kcal/mol (based on a K_i of 0.84 nM for grazoprevir against the R155K HCV NS3/4a protease mutant[23]).

Our prior work focusing on macrocyclic peptide strain also included data from non-peptidic macrocycles and from non-macrocyclic small molecules.[21] We observed a rough upper bound on ligand strain, independent of ligand type and linear in molecular size, of approximately $0.3(N_{nonH} - 10)$ kcal/mol (for grazoprevir, with 54 non-hydrogen atoms, the nominal upper bound is 13.2 kcal/mol). Here, we show that the distribution of strain energy follows a normal distribution (with an enforced lower bound of 0) whose mean and variance can be defined as a function of molecular conformational complexity that is superlinear with respect to molecular size. The distributional model fits strain calculations for nearly 3,000 PDBBind[24,25] protein−ligand complexes whose ligands were subjected to real-space refinement with xGen (in addition to the prior data set of 341 protein−ligand complexes that was skewed heavily toward large macrocycles).

Further, the relationship between calculated strain and experimentally determined binding affinities comports with physical principles. There is a quantitative relationship between ligand efficiency (binding free energy per non-hydrogen atom) and ligand strain-per-atom. Ligands with high efficiency seldom have high per-atom bound conformational strain, but ligands with low efficiency exhibit a broad range of per-atom strain.

The distributional model, in the context of the relationship between strain and binding affinity, has direct implications for conformational search protocols as well as for molecular design and optimization.

Data and methods discussed in this paper are available to other researchers (see Data Availability Statement).

RESULTS AND DISCUSSION

We introduce a distributional model of bound ligand strain that accounts for the dependence on molecular size using a macrocycle-focused set of 341 protein−ligand complexes[21] (the "Brueckner Set") augmented with nearly 3,000 complexes from the PDBBind 2020 refined database[24,25] (the "PDBBind Set"). After describing the data set characteristics, we summarize the differences between the standard approach to modeling ligand strain and the xGen methodology. Next, we illustrate the form of the ligand strain distributional model using the macrocycle-focused data set and present refinements to the model using the PDBBind protein−ligand complexes. The relationship to alternative approaches and the biological relevance of an accurate means for estimating and modeling strain are discussed last.

Molecular Data Set Composition. The Brueckner Set was comprised of 38 molecules from the Perola/Charifson study[3] for which electron density was available, 147 non-peptidic macrocycles,[20] and 156 peptidic macrocycles.[21] These ligands were both larger and more flexible than typical drug-like molecules. In order to understand bound ligand strain more generally, here we have performed real-space ligand refinement on the full PDBBind refined data set (2020 release),[24,25] resulting in 2,996 cases having high-quality ligand conformational ensembles, each with annotated experimentally determined binding affinities (see Experimental Section for details). Figure 2 summarizes the size, flexibility, and binding site contact for the ligands in the Brueckner Set and the PDBBind Set.

The PDBBind ligands generally exhibit the size and complexity expected of drug-like molecules, per the curation goals of the refined set.[25] As shown in Figure 2, roughly 90% of the ligands have 40 or fewer non-hydrogen atoms (roughly less than 500 Da in molecular weight) and 80% have 10 or fewer rotatable bonds. In contrast, the Brueckner Set has roughly 70% of ligands with *more* than 40 non-hydrogen atoms and 90% with *greater* than 10 rotatable bonds.

Molecular size is also related to the proportion of ligand atoms that are in close contact with a protein when bound. A very simple measurement of the degree of contact is the proportion of ligand atoms (including hydrogens) whose van der Waals (VdW) surfaces are within 1.0 Å of the nearest protein atom's VdW surface. Figure 2C shows a clearly multi-modal distribution of ligand contact for the macrocycle-focused set. A threshold of 70% contact roughly splits the Brueckner Set between distributional peaks. The PDBBind Set has a small tail of ligands with contact lower than 70% (less than 1% of cases). As seen in Figure 2D, very large ligands (roughly 80 non-hydrogen atoms or more) tend to skew toward lower fractional contact. However, smaller ligands span a range of binding site contact. For example, the 40−60 atom size range includes ligands ranging from 35% to 100% binding site contact. For reference, the macrocycle in Figure 1 has 82% binding site contact and 54 non-hydrogen atoms. In what follows, the subset of the combined Brueckner and PDBBind Sets with at least 70% contact (the "High-Contact Set") will be analyzed in the greatest detail, as it comprises nearly all of the data (3,148 of 3,337 cases).

To provide context regarding the diversity of ligands and binding site contact, Figure 3 shows six representative examples. N specifies the number of non-hydrogen atoms, energy values are in kcal/mol, and ΔG values were estimated from PDBBind unless otherwise noted. Minimum RMSD values are given for the xGen ensembles compared with the PDB ligand coordinates. The examples span the small but highly potent case of biotin bound to a streptavidin mutant (Panel A) to very large peptide macrocycles with divergent binding site contact (Panels E and F). The xGen approach yielded strain reductions in all cases, and the conformer ensembles showed conformational heterogeneity in solvent-exposed areas. Each also included a conformer that was quite close to the original PDB ligand coordinates.

Real-Space Ligand Refinement: Low-Energy Conformational Ensembles. As shown in Figure 1, modeling ligand density may be done with a conformational ensemble rather than a single conformer, as previously described in detail.[20,21] Briefly, a conformational search is carried out in which each conformer's fit to real-space electron density is expressed as an energetic reward, balanced against the energy from a variant of MMFF94s. Figure 4 shows the resulting conformer set for the 3SUE case introduced above (top center, shown with a heat-map of real-space electron density). These conformers balance the tension of fitting the density perfectly against the constraints of conformational energy.

Each of these balanced-pool conformers is then re-minimized in two ways: a) under a condition in which the density overlap is strongly weighted (top right) and b) with no density overlap reward but with a positional restraint (top left). These three pools are used to identify conformer trios with the property that a low-energy "min-pool" member and a high-fit-quality "density-pool" member are both within a small neighborhood of a central "balanced-pool" member. The three members of such a trio are shown at the bottom of Figure 4 along with their associated energy values. The collection of such trios is then used to identify a conformer ensemble to optimize the fit to real-space

https://doi.org/10.1021/acs.jmedchem.2c01744
J. Med. Chem. 2023, 66, 1955−1971

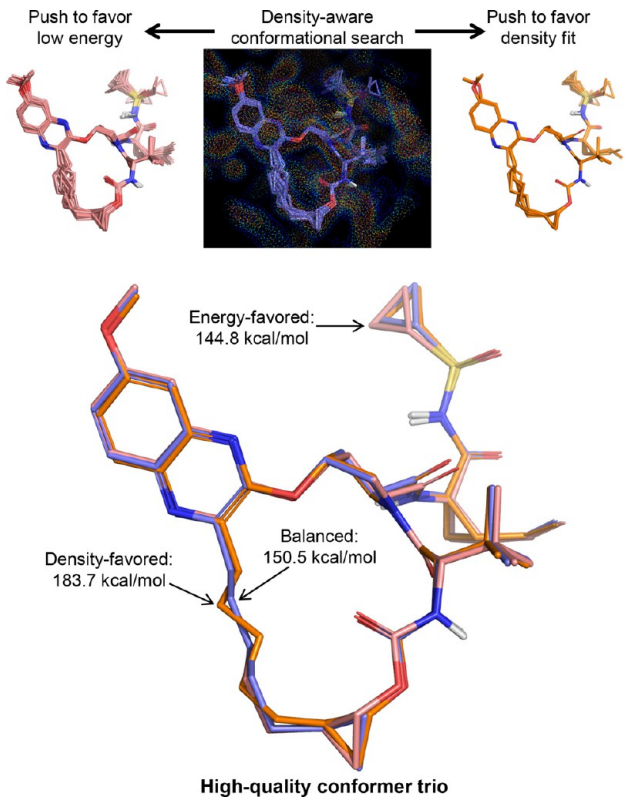


Figure 4. The xGen approach constructs ensembles to fit electron density from conformational trios, each of which is anchored by a conformer that arose from an X-ray density-aware conformational search (slate). These balanced conformers are optimized toward a density-weighted fit (orange) and an energy-weighted fit (salmon). Conformer trios are identified whose density-favored member and energy-favored member are near optimal and where both are geometrically close to a central balanced conformer.

electron density. Note that the density-pool members are used to construct the occupancy-weighted conformer ensemble, and their corresponding min-pool siblings serve as energy-surrogate conformers.

The respective conformational energy values for the three members of the trio in Figure 4 were 183.7 (density-favored), 150.5 (balanced), and 144.8 (energy-favored). This particular trio had the single best-fitting density-favored conformer, and that excellent fit to experimental X-ray data is reflected by very high conformational energy. Position-restrained relaxation of that conformation improves its energy (to 157.0 kcal/mol) with minor atomic coordinate movement (0.1 Å RMSD). However, that is still much higher energy than seen with the energy-favored sibling conformer within the trio, which is *not reachable* through local optimization. The phenomenon of nearby local energetic minima that are unreachable is the principal problem with the standard approach to identifying energy-surrogate conformers.

By constructing the conformational ensemble with an awareness of the importance of energy-surrogate conformers, xGen obtains high-quality fits to electron density while also allowing for estimation of energetically reasonable bound ligand strain. The difference between the minimum energy obtained from the set of energy-surrogate conformers and the global energy minimum for the ligand is the reported strain energy (the Experimental Section contains additional details). In the case shown in Figures 1 and 4, the net result is a marginally better fit to electron density with a four-conformer ensemble along with a 75% reduction in estimated strain energy.

For the PDBBind Set, the average ensemble size was 2.5, with a small but consistent improvement in real-space density fit and an average reduction in estimated strain energy of 37% compared with the standard approach of positionally restrained minimization of the original deposited ligand coordinates. The mean deviation for the final energy-surrogate conformers to the

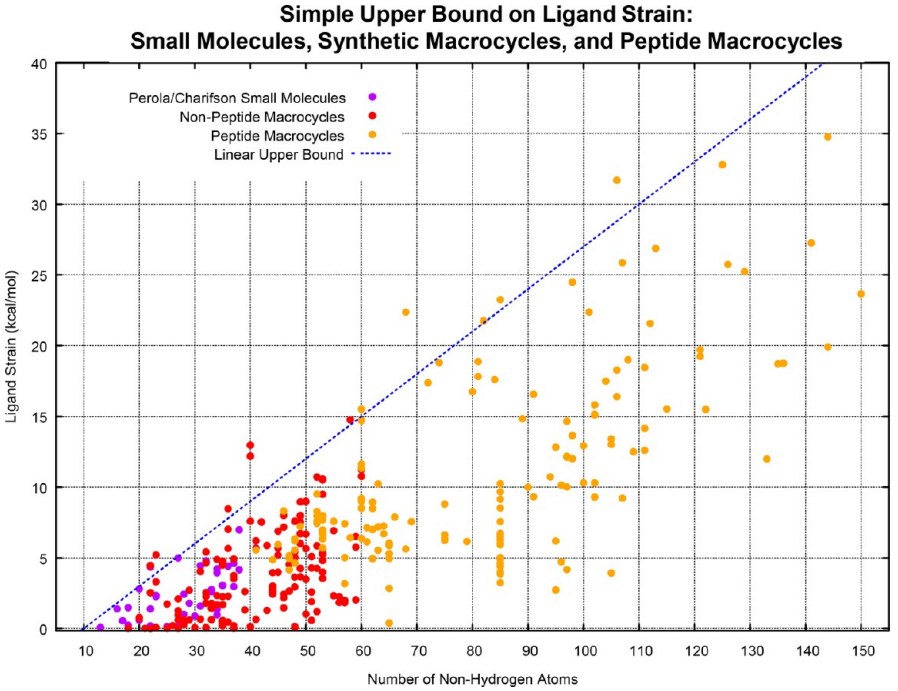


Figure 5. Relationship between strain energy and molecular size for the Brueckner Set with an approximately linear upper bound on strain (blue dashed line).

https://doi.org/10.1021/acs.jmedchem.2c01744
J. Med. Chem. 2023, 66, 1955−1971

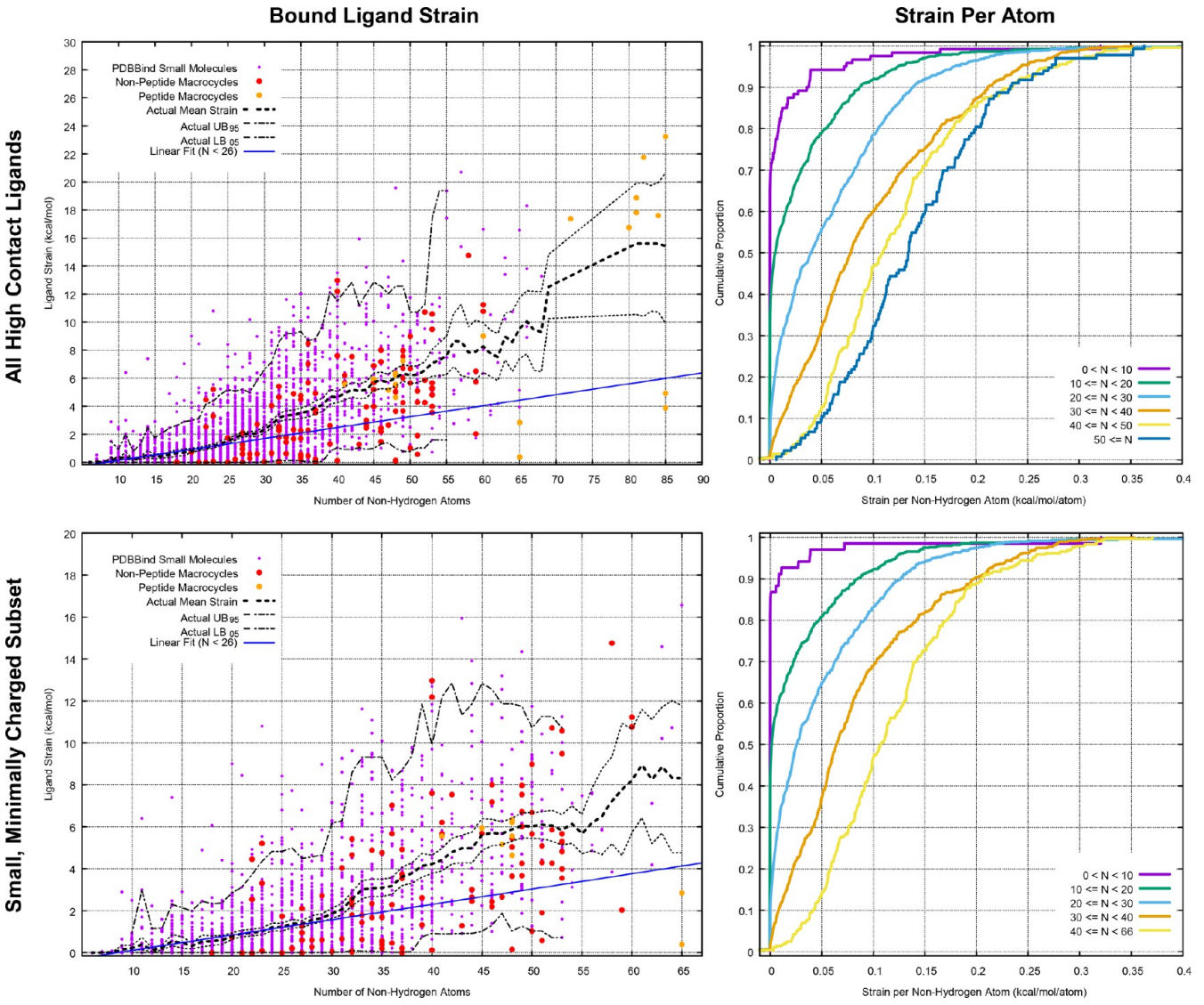


Figure 6. Relationship of ligand strain to molecular size.

real-space fitted conformers was 0.3 ± 0.12 Å on average. For the Brueckner Set, the average ensemble size was 3.5, again with a small but consistent improvement in real-space density fit and an average reduction in estimated strain energy of 43% compared with the standard approach. The mean deviation for the final energy-surrogate conformers was 0.3 ± 0.14 Å on average for the Brueckner Set.

The primary focus of this study is strain energy, not the comparative advantages of ensemble-based ligand modeling in X-ray diffraction data. However, for this study, it is important that the ensemble approach yielded ligand models of similar quality to the original deposited ligand coordinates. As seen in our prior work,[20,21] improvements in RSCC and RSR were small but consistent compared with original PDB ligand coordinates: an average RSCC of 0.924 and an average RSR of 0.156, compared to the original PDB ligand coordinate metrics of 0.897 and 0.169, respectively.

Ligand Strain Is Superlinear in Ligand Size. Figure 5 shows a simple linear upper bound suggested by the derived ligand-strain values (S_{UB} = 0.3(N_{nonH} − 10) kcal/mol), as previously described.[21] While this apparent bound accounts for the lower-right triangular relationship between ligand strain and molecular size, it lacks statistical support, and it does not account for the apparent lack of very low strain values at the higher range of molecular size.

Considering the distribution of strain normalized by N_{nonH}, the Perola/Charifson subset (violet in Figure 5) and the non-peptide macrocycles (red) both shared a median value of 0.08 kcal/mol/atom. However, the peptide−macrocycle subset (orange) was more than 60% larger: 0.13 kcal/mol/atom. This observation led to the hypothesis that the expectation value for ligand strain might be superlinear. Given the limited number of data points, especially at larger molecular sizes, adding the PDBBind Set to the Brueckner Set provided an opportunity to better quantify the relationship between ligand size and strain.

Figure 6 shows the relationship between estimated global strain for the High-Contact Set (top two plots, 3,148 total protein−ligand complexes) and for a simplified subset with one formally charged atom or fewer and 65 non-hydrogen atoms or fewer (bottom two plots, 2,282 protein−ligand complexes). The left-hand plots illustrate the relationship between ligand strain and molecular size, with the PDBBind cases shown in violet, non-peptide macrocycles shown in red, and peptidic macrocycles shown in orange. Smoothed sample mean strain is plotted

https://doi.org/10.1021/acs.jmedchem.2c01744
J. Med. Chem. 2023, 66, 1955−1971

with a thick dashed line (with its confidence interval shown in thin dashes), smoothed sample upper and lower 95th and 5th percentiles are plotted with dashes and dots, and the linear fit to the strain for small molecular sizes is plotted with a thick blue line. The right-hand plots show cumulative histograms of strain-per-atom in increasing size bins.

For the full High-Contact Set, ligand strain is clearly superlinear: the linear fit to strain clearly diverges as molecular size increases (top-left plot of Figure 6), and there are consistent rightward shifts in cumulative histograms of strain-per-atom as molecular size increases (top-right plot of Figure 6). However, the full High-Contact Set contains cases with a number of features that complicate accurate strain estimation and may lead to erroneously high strain estimates: additional degrees of freedom that might hamper identification of low-energy ensembles, salt-bridge collapse in cases with multiple formal charges leading to artificially low global minimima, the effects of interstitial water within large ligands that may affect structural stability, and cases where multiple proximal like-charged moieties, treated as ionized, may lead to high estimates of bound conformational energy.

The simplified subset contains over 2,000 cases, even with restrictions on molecular size (65 or fewer non-hydrogen atoms) and formal charge (one or fewer ionizable ligand atoms). With these restrictions, the concerns about overestimated strain energies were ameliorated. As seen in Figure 6 (bottom two plots), elimination of the potentially problematic cases had essentially no effect on the observations regarding superlinearity.

Note that others have attributed anomalously high strain to artifactual polar collapse,[28] which could account for the observed superlinearity. However, the simplified subset eliminated the most potentially problematic cases, and the superlinearity remained. We also recalculated global strain values for the entire data set of over 3,000 cases *without* a polar component in the energetic calculations. Results were nearly identical, with per-ligand global strain differing by 0.32 kcal/mol on average between the xGen-based global strain calculation and this apolar control calculation, with $\tau = 0.92$ ($p \ll 0.001$, Pearson's $r = 0.99$).

It is not clear how to directly confirm the superlinearity experimentally, but there are clear physical principles that support the observation. There are three components to global ligand strain: 1) the X-ray fitting strain, which is generally considered to be an artifact of crystallographic data reconstruction, and which is addressed by xGen ligand refinement and positionally restrained minimization; 2) the local strain, which is the difference between the positionally restrained conformer and an unrestrained local minimum; and 3) the difference between the energy of the unrestrained local minimum conformer and the global minimum.

One might expect that the local strain would be proportional to molecular size, given that, under reasonable assumptions of gross molecular shape (e.g., cylindrical as ligands become large), the binding contact area and maximal interaction energy increase roughly linearly with size. Consequently, the distortive effect on the ligand would also be roughly linear. However, the difference in energy between the unrestrained local minimum conformer and the global minimum should *also* be related to conformational flexibility: the likelihood that a ligand might adopt a much lower energy conformation in its unbound state should increase with ligand flexibility.

Within the high-contact combined set, molecular size bins of 10−15, 16−20, 21−25, 26−30, 31−35, and 36−40 all had at least 200 cases. For each of those narrowly defined ligand size bins, there was a statistically significant positive correlation ($p < 0.001$, by Kendall's Tau) between number of rotatable bonds and ligand strain. Conversely, for rotatable bond counts of 0−1, 2−3, ..., 10−11, and 12−15 (each with at least 200 cases), there was a statistically significant positive correlation ($p < 0.001$, by Kendall's Tau) between number of non-hydrogen atoms and ligand strain for all flexibility bins.

Given the clear superlinear relationship between molecular size and ligand strain, a simple proposition is that the relationship might be roughly quadratic in terms of molecular size. A quadratic relationship offers three interesting physical underpinnings. First, as molecular size increases, non-bonded energy terms, which increase with the square of the number of atoms in a molecule, begin to dominate in influence. Second, demonstrated nicely by Reynolds et al.,[29] as size and complexity increase across multiple specific protein−ligand interactions, structural compromises in the form of induced strain and suboptimal protein−ligand complementarity multiply, contributing to increasing per-atom strain. Third, the square of the number of non-hydrogen atoms is strongly correlated with the product of molecular size (the number of non-hydrogen atoms) and molecular flexibility (the total number of exocyclic and macrocyclic rotatable bonds plus one). For the High-Contact Set, Kendall's Tau was 0.76 ($p \ll 0.001$), and Pearson's r was 0.94 (see Figure 7).

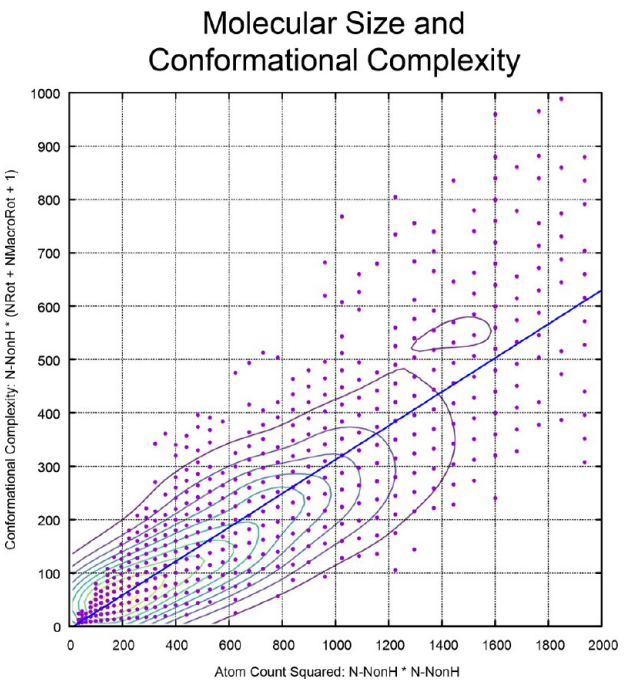


Figure 7. Relationship between the square of heavy atom count (X axis) and conformational complexity modeled as the product of atom count and number of rotatable bonds (Y axis).

A Distributional Model of Ligand Strain. The strain for a ligand with N non-hydrogen atoms appears to behave like a random variable x that follows the distribution defined by f:

$$f(x;\, \mu,\, \sigma^2) = \begin{cases} \dfrac{1}{\sqrt{2\pi\sigma^2}}\, e^{(x-\mu)^2/2\sigma^2}, & \text{if } x > 0 \\ 0, & \text{otherwise} \end{cases} \quad (1)$$

https://doi.org/10.1021/acs.jmedchem.2c01744
J. Med. Chem. 2023, 66, 1955−1971

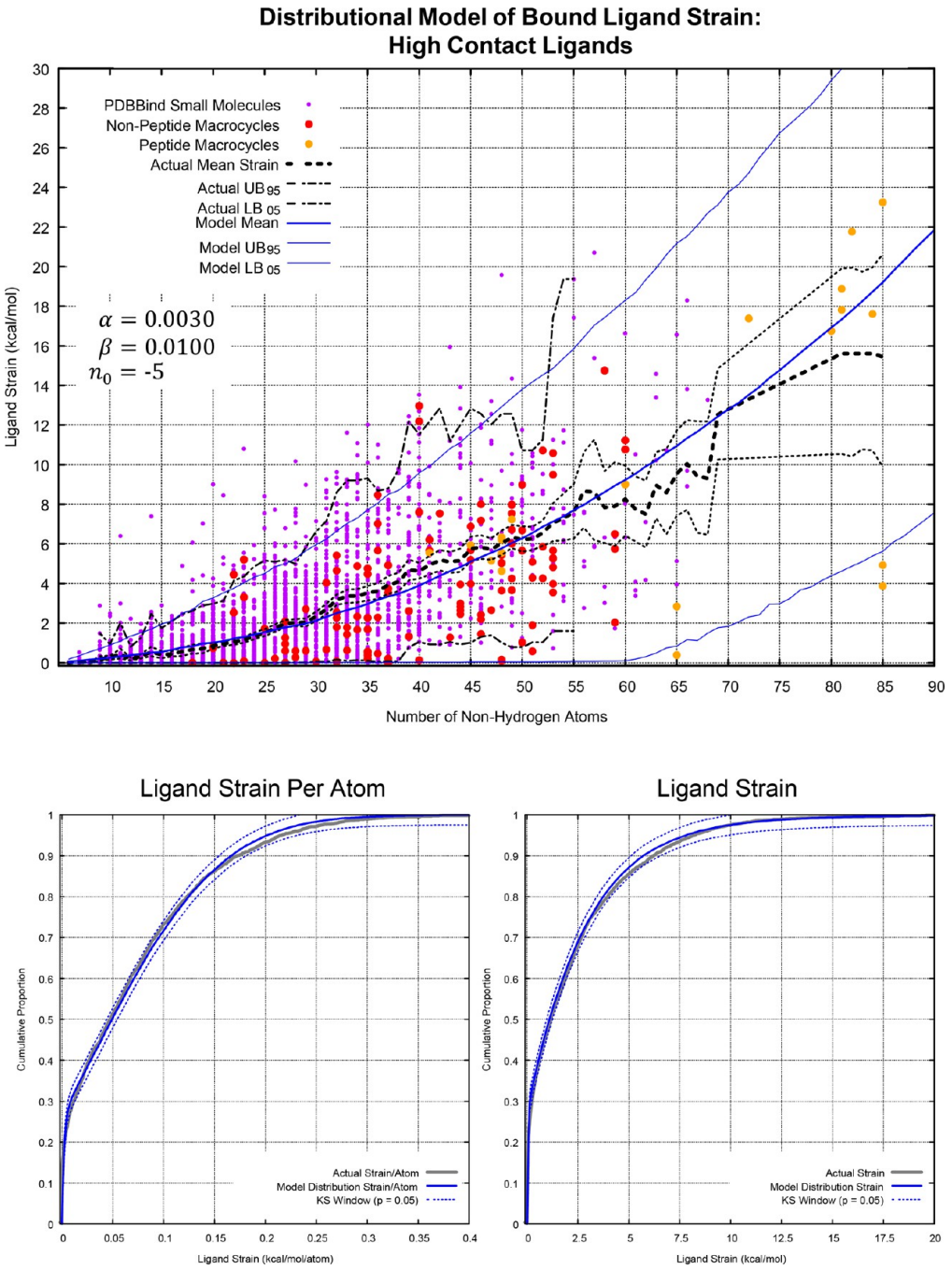


Figure 8. Relationship between strain energy and molecular size for the High-Contact Set, modeled using the proposed size-dependent rectified normal distribution.

$$\mu = \alpha(N + n_0)^2 \quad (2)$$

$$\sigma^2 = \beta(N + n_0)^2 \quad (3)$$

where α, β, and n_0 are empirically determined constants and, as will be shown below, have defined but different set of values for high- vs low-contact ligands. Note that f is a normal distribution with an enforced lower bound of 0 (also known as a rectified normal distribution) whose mean and variance are quadratically dependent on molecular size.

Figure 8 illustrates the distributional model for the High-Contact Set. The top plot is as shown in Figure 6, with the addition of the distributional-model-predicted mean and upper and lower bounds (blue curves). Using the exact molecular sizes seen in the set, predicted strain energy values were synthetically generated from the model distribution using the μ and σ predicted for each molecular size. Note that the smaller molecular size range (up to 40 non-hydrogen atoms) contains nearly 3,000 data points, and the sample-mean estimate has a correspondingly narrow confidence limit that follows very closely to the modeled distribution. As data density decreases with molecular size, the confidence limit for the sample mean widens, but it still largely encloses the modeled distribution mean. The upper and lower bounds were estimated from the observed and synthetic distributions at the 95th and 5th percentiles, respectively. For the observed data, the bounds were relatively noisy, as expected from sampled tail data.

https://doi.org/10.1021/acs.jmedchem.2c01744
J. Med. Chem. 2023, 66, 1955−1971

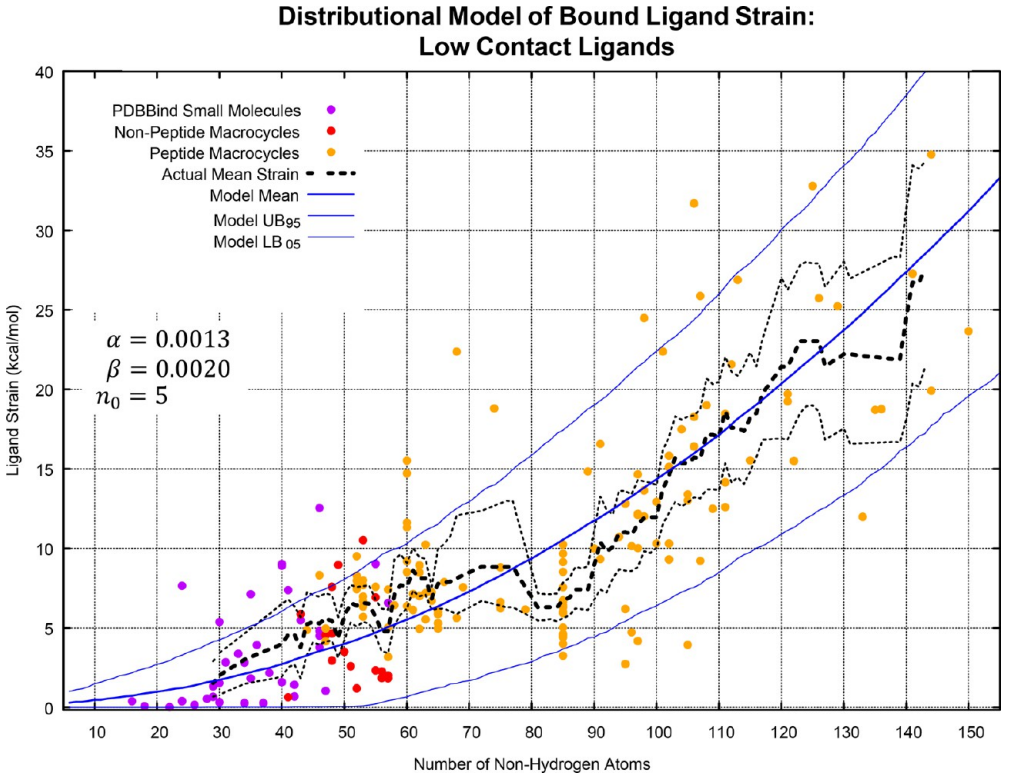


Figure 9. Relationship between strain energy and molecular size for the Low-Contact Set, modeled using the proposed size-dependent rectified normal distribution.

The bottom two plots of Figure 8 show the comparison of the cumulative histograms of observed ligand strain with the synthetic strain values (gray and blue solid curves, respectively), both for ligand strain normalized per non-hydrogen atom (left) and for overall ligand strain (right). If our model distribution was a poor model of the actual distribution, then the cumulative histograms would diverge substantially. The dashed blue lines form a band whose width is determined using the Kolmogorov–Smirnov (KS) test statistic. Given that the gray curves are contained within the bands, one cannot say that the observed and synthetic distributions are different from one another with any degree of confidence, even at the weak p-value of 0.05. The actual sample distribution is therefore likely to be well-modeled by the proposed rectified normal distribution.

The KS test statistic is constructed to reject the hypothesis that two distributions are the same. Therefore, small p-values require larger deviations between two distributions in order to support a claim that the distributions are different. Here, we are illustrating the *closeness* of our distributional model and the strain energy estimates. So when the experimental distribution falls within the bands around the model distribution, the higher the nominal p-value is, the more likely it is that the distributions are the same. Note also that larger data set sizes lead to narrower bands. In Figure 8, the bands are narrow, given the large data set (>3,000 cases) coupled with the high nominal p-value (= 0.05), highlighting the tight correspondence between the observed and synthetic model distributions.

The Low-Contact Set (less than 70% binding site contact) includes just 189 ligands in total. Figure 9 is analogous to the top plot seen in Figure 8 but for the Low-Contact Set. The sample mean estimate was quite noisy, due to the sparsity of data, but tracked well with the modeled distribution (sample upper and lower bounds could not be estimated). As with the High-Contact Set, the observed strain-per-atom and total strain fell within conservative bands around the modeled distribution (data not shown, with the bands being substantially wider due to the much smaller data set).

Distributional Model Variants. We have shown that a simple distributional model, with just three parameters, fits the observed bound ligand strain for a very broad variety of ligands using a quadratic dependence on molecular size, measured as the number of non-hydrogen atoms.

It is also possible to model the distribution as being *linear* with respect to mean and variance using the product of the number of non-hydrogen atoms and the total number of freely rotatable bonds plus one (including both exocyclic and macrocyclic bonds) as a simple conformational complexity measure. A linear model exhibits roughly the same quality of fit in terms of the distribution of observed to modeled ligand strain as seen for the quadratic model, because of the strong correlation between the square of the atom count and the conformational complexity measure (see Supporting Information for details). However, the quadratic model is simpler, requiring only a count of the number of non-hydrogen atoms. Also, it directly produces an estimate for strain-per-atom, while the linear model cannot do so, being parameterized on the conformational complexity measure.

Ligands with high binding site contact are the overwhelming majority of drug-like ligands, and their strain was well modeled with the quadratic molecular-size approach using the rectified normal distribution defined in eqs 1–3 with α = 0.0030, β = 0.0100, and n_0 = −5. Ligands with low binding site contact, typified by large peptidic macrocycles but including smaller macrocycles and some "normal" small molecules, were well modeled with α = 0.0013, β = 0.0020, and n_0 = +5. The major difference between the high- and low-contact models is that, for

https://doi.org/10.1021/acs.jmedchem.2c01744
J. Med. Chem. 2023, 66, 1955−1971

a given molecular size, the high-contact model predicts higher strain than the low-contact model.

The analytical formulas for the 95th percentile upper and 5th percentile lower bounds on strain are

$$UB_{95} = \mu + 1.645\sigma \quad (4)$$

$$LB_{05} = \max(\mu - 1.645\sigma, 0) \quad (5)$$

where μ and σ are defined in eqs 2 and 3. Note that the true lower bound, while close to eq 5, is very slightly shifted upward near the intersection with 0 strain energy due to the effect of clipping in the rectified normal distribution.

Given a ligand with 30 non-hydrogen atoms, the high-contact model predicts a mean strain of 1.9 kcal/mol with a 95th percentile upper bound of 6.0 kcal/mol. Respectively, the low-contact model predicts 1.6 and 4.2 kcal/mol. The trend of lower predicted strain values for the low-contact condition increases in magnitude as ligand size increases. At 60 non-hydrogen atoms, the expected mean/upper bound shifts to 5.5/10.3 and 9.1/18.1 kcal/mol for the low- and high-contact models. This seems to be an intuitive result: with higher contact, greater free energy gains can accommodate higher levels of strain. As the influence of the binding event decreases, with lower contact and reduced free energy gain, the impact on the conformational preference of the ligand should also decrease. In the asymptotic case of vanishing binding site contact, we would, of course, expect to see no strain at all.

It seems surprising that the objects of a human design process should have a key property that follows a relatively simple distribution at all. A number of molecular properties are actively tuned during lead optimization, for example: potency, selectivity, logP, size, formal charge, pharmacokinetics, and metabolic liabilities. Of these, only potency and selectivity have direct links with bound ligand strain, and considerations of strain sometimes do influence molecular design choices. However, because strain cannot be measured experimentally and computational strain modeling has been an evolving area, observed bound ligand strain, to a degree, is an experiment of nature. To the extent that direct optimization of ligand strain has been (or will be) undertaken for particular targets or within particular molecular series, the distributional model derived here will begin to break down.

Relationship to Other Work. In a very early study of bound ligand conformational strain energy using the CHARMm force field, Nicklaus et al.[1] found relatively high average conformational strain energies (15.9 kcal/mol) across a set of 27 flexible ligands. Subsequently, Boström et al.,[2] using MM3 and AMBER, estimated quite low energies for 70% of studied ligands (≤3 kcal/mol) but found a significant set of nominal outliers with higher energies (roughly 20% with strain of 6−19 kcal/mol). Perola and Charifson[3] characterized a 150 ligand set (very large for the time) using MMFF and OPLS-AA, and found roughly 60% of cases to have strain energy ≤5 kcal/mol but at least 10% to have strain energy ≥9 kcal/mol. Each of these studies used minor variations of what we have termed the "standard" method for strain energy estimation, and all focused on small molecule ligands.

By contrast, for the high-contact cases in the combined Brueckner and PDBBind Sets with 40 or fewer non-hydrogen atoms (2,797 complexes in total), 70% of the xGen refined ligands had strain values of ≤2.1 kcal/mol, with just 10% ≥4.6 kcal/mol, and less than 2% ≥9.0 kcal/mol. Including the larger high-contact ligands (3,148 cases in total), the values shifted slightly, to ≤2.6 kcal/mol at 70%, ≥6.2 kcal/mol at 10%, and ≥9.0 kcal/mol at less than 4%.

Updating the prior study of Nicklaus et al.,[1] Sitzmann et al.[7] calculated conformational energies at the DFT level of theory. They carefully curated high-quality PDB structures using criteria about structure quality (e.g., resolution, R_{free}, and real-space ligand fit), ligand appropriateness (e.g., non-ionizable molecules with limited conformational flexibility), and numerous other aspects, resulting in 98 ligand instances in their strictest filtered subset. Nominal calculated conformational energies were much higher than reported here, with a median value of 9.2 kcal/mol and 70% having energies greater than 6.4 kcal/mol, compared with results here showing strain energies ≤2.1 kcal/mol for 70% of cases in the combined Brueckner and PDBBind Sets with 40 or fewer non-hydrogen atoms.

More recent work from Zivanovic et al.[10] considered 115 cases (with substantial overlap to the Perola/Charifson set) using high-level QM methods for energy evaluation and unconstrained minimization of the bound ligand conformations to arrive at energy-surrogate conformers. The latter choice led to greater divergence from crystallographic ligand coordinates than is typical in such studies, allowing for lower bound strain estimates than more stringent protocols for managing crystallographic deviation. Nonetheless, their results largely paralleled our own, with 73% of ligands yielding strain ≤1.8 kcal/mol (compared with 66% observed here in the combined Brueckner and PDBBind Sets with 40 or fewer non-hydrogen atoms).

Another, larger-scale, QM approach is the recent study by Tong and Zhao,[9] which considered over 6,000 cases of protein−ligand complexes using *ab initio* energy calculations and a relatively stringent approach to energy-surrogate conformer derivation. They filtered publicly available structures for structural quality in addition to restricting molecular complexity (130 < MW < 600 and number of rotatable bonds ≤10). Overall, they obtained a mean strain energy of 4.6 kcal/mol (median of 3.7 kcal/mol) and a 95th percentile of 12.4 kcal/mol. The set of cases included 828 in common with the combined Brueckner and PDBBind Sets studied here, in which the respective *ab initio* strain energy values were comparable to the full set: 4.7 (mean), 3.6 (median), and 12.6 kcal/mol (95th percentile). The xGen-based strain energy estimates were much lower: 1.5 (mean), 0.8 (median), and 5.8 kcal/mol (95th percentile).

Non-aromatic rings were identified as often yielding anomalously high strain energy,[9] and the example of PDB code 4MU7 was highlighted, with a calculated *ab initio* strain energy of 21.7 kcal/mol for the ligand. Figure 10 shows the deposited and xGen refined ligand models, with the former yielding 12.1 kcal/mol in strain energy and the latter yielding 1.1 kcal/mol (both values calculated as above for the Brueckner and PDBBind data sets). Re-refinement of the ligand produced an obviously improved structure, with a believable estimated strain energy.

High-Quality Ligand Fitting Is Critical. In estimating bound ligand strain energies, theoretical sophistication in calculations of energy values cannot overcome poorly fit ligand models. Poor fit, in the sense being used here, may not be detected by metrics such as RSCC, RSR, or EDIAm.[22,30] The nominal fit quality for the deposited ligand model in Figure 10 (PDB code 4MU7) was high using all three metrics. Cases of poor fit, as in this example, will generally produce erroneously high strain estimates. When low-energy models exist that fit the X-ray data as well or better, strain estimates will more closely track with measured binding affinity values (i.e., by not

https://doi.org/10.1021/acs.jmedchem.2c01744
J. Med. Chem. 2023, 66, 1955−1971

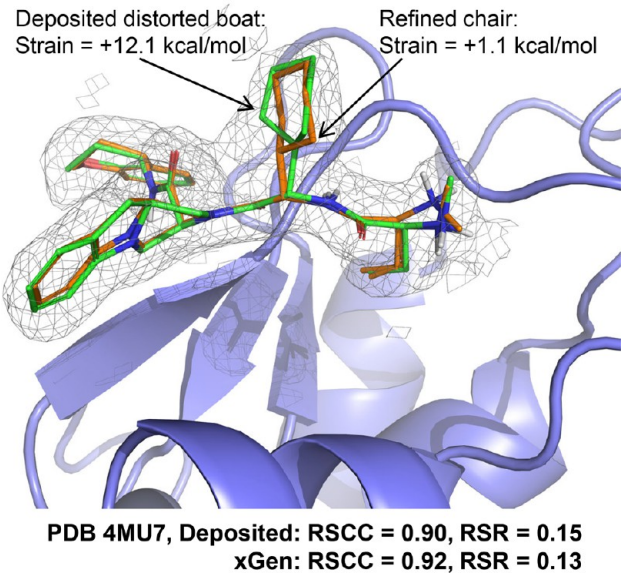


Figure 10. Typical example of strained cyclohexane conformation contributing to high nominal strain in an uncorrected PDB structure (green) being resolved by re-refinement (orange).

exceeding reasonable expectations based on the free energy of binding).

Fit quality, absent consideration of ligand energetics, cannot be used to identify cases where crystallographic ligand models may be used as starting points for strain estimation, whether using simple force field methods or highly sophisticated *ab initio* methods. Cases of poor fit, exemplified by the ligand of 4MU7, will tend to produce anomalously high strain estimates. Even in cases where the quality of ligand models is not an issue, computationally expensive calculations are difficult to apply rigorously to obtain reliable global minimum estimates, effectively adding another source of noise to strain estimates.

Alternate Energetic Calculations. As we have just seen, strain estimation must begin with energetically reasonable ligand conformations. The foregoing xGen-based results relied upon a variant of MMFF94s, which is limited, to a degree, in detailed parameterization of different torsion types. Some studies of ligand strain have made use of multiple alternative force fields,[2,3] finding relatively little difference in general trends or estimated strain bounds. Here, the xGen conformer ensembles fit to X-ray density are strongly influenced by the experimental diffraction data. The energetic benefit of correctly overlapping a single carbon atom with matching density is roughly −8.0 kcal/mol,[20] which is enough to overcome inaccuracies in torsional parameterization. So, the geometries of the conformer ensembles are well-fit to the experimental X-ray data while avoiding obviously high energy configurations.

However, when making quantitative estimates of strain using the energy-surrogate conformer from the xGen trio, inaccuracies in torsional parameterization may become important. In order to address this, we made use of OPLS4[31] for a randomly selected 10% of the PDBBind Set (see Experimental Section for details). The correlation between the xGen strain estimates and the OPLS4 estimates was very high: $\tau = 0.59$ ($p \ll 0.001$, Pearson's $r = 0.78$). The OPLS4 *post hoc* strain estimates derived from the xGen conformer ensembles increased compared with the original xGen estimates by a median of 0.9 kcal/mol (mean strain increased by 1.4 kcal/mol).

We also calculated OPLS4 strain estimates associated with the deposited PDB coordinates using the same procedure. OPLS4 strain derived from the xGen ensembles was consistently lower than was derived from the PDB coordinates, with strain decreasing by a median value of 1.2 kcal/mol (mean strain decreased by 2.2 kcal/mol). In 116/299 cases (39%), strain was reduced by 1.0 kcal/mol or more, and in just 5/299 (2%) of cases was strain increased by 1.0 kcal/mol or more. Typical strain reduction for OPLS4 was roughly 40%, which was similar in magnitude to that seen with the results presented above. The process of energy-aware real-space refinement of ligand conformer ensembles substantially reduced estimated strain compared with deposited PDB coordinates using either force field.

We have not carried out higher-level energy calculations, for two reasons. First, while obtaining reasonable conformational energy estimates for the bound state of ligands is feasible, obtaining reliable estimates for global minima is very challenging, especially for complex molecules. Second, as a practical matter, force field methods can be applied in lead optimization scenarios, even on macrocyclic ligands. As will be seen next, the xGen-based force field estimates of strain are directly related to experimentally measured binding affinities, suggesting that, even given limitations in the accuracy of the strain estimates, they are directly relevant to the molecular design problem.

Biological Relevance. The foregoing has demonstrated that the xGen method identifies lower energy bound ligand conformations than other approaches and that the distribution of observed strain energies closely follows a simple molecular-size-dependent distribution. However, to this point, we have made no connection between experimental measurements and calculated strain energies to support the idea that they reflect actual physically meaningful phenomena. As mentioned earlier, it is not clear that a *direct* experimental measurement can be made in support of the quality of the strain estimates. A key advantage of the carefully curated PDBBind Set is that each protein−ligand complex is associated with an experimentally determined affinity or activity, expressed as a K_d, K_i, or IC_{50}, in decreasing order of prevalence. Each of these can be converted into an approximate free energy of binding, and then into a measurement of ligand efficiency (LEFF) that captures the free energy gained per non-hydrogen atom in a given ligand. This offers an *indirect* means to assess the quality of strain estimates.

The relevance of LEFF to ligand strain is that, given a highly efficient ligand, one expects to see low (or no) strain. However, given an inefficient ligand, it may lack efficiency due to high strain, but it may also lack efficiency based on poor complementarity to its binding site (with little or no strain). Figure 11 shows the relationship between molecular size and LEFF along with the relationship between LEFF and ligand strain-per-atom. The top plot recapitulates the observations of Kuntz et al.[32] in one of the foundational papers introducing the idea of ligand efficiency. As molecular size increases, the upper bound on LEFF decreases (for the PDBBind data, direct correlation measured by Kendall's Tau (τ) was −0.55, $p \ll 0.001$).

Many reasons have been proposed for this strikingly consistent relationship,[33] but the distributional model we have proposed here offers an obvious contributor: ligand strain increases with molecular size (importantly, even on a per-atom basis) due to its quadratic relationship with the modeled strain expectation value and upper bound. Figure 11 (middle plot)

https://doi.org/10.1021/acs.jmedchem.2c01744
J. Med. Chem. 2023, 66, 1955−1971

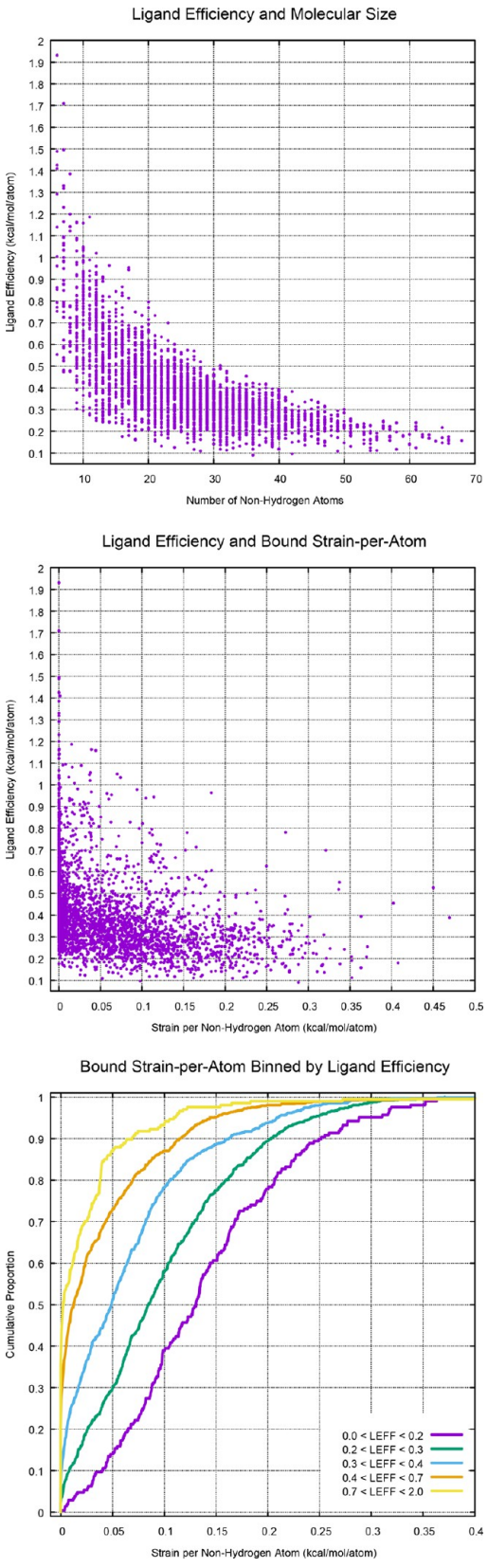


Figure 11. Relationship of ligand efficiency to ligand size and ligand strain-per-atom.

shows that the relationship between strain-per-atom and LEFF is as expected. In cases with high LEFF, few molecules exhibited high strain-per-atom. In cases with low LEFF, strain-per-atom was spread across a broad range. The inverse correlation between LEFF and strain-per-atom was highly statistically significant ($\tau = -0.35$, $p \ll 0.001$, 95% CI: −0.37 to −0.33). Note that if ligand strain were linear instead of superlinear, this correlation would disappear.

These trends are more easily seen by binning ligand efficiency and considering cumulative histograms (bottom plot of Figure 11). The highest and lowest efficiency bins had roughly 200 data points, with the central bins each having roughly 800−900. In the highest efficiency bin (yellow curve), nearly 90% of molecules had strain-per-atom ≤0.05 kcal/mol/atom. This fraction dropped with each step down in efficiency, to just 15% in the lowest efficiency bin, where the strain-per-atom is spread widely (violet curve).

The quantitative relationship between estimates of ligand strain and experimentally measured ligand efficiency offers a means to assess the quality of strain estimates. The correlation between the standard strain estimate based on deposited crystallographic coordinates was significantly lower ($\tau = -0.27$, $p \ll 0.001$, 95% CI: −0.30 to −0.26) than for the xGen-ensemble estimates. Of note, the *ab initio* strain estimates for the 828 PDBBind ligands from the Tong and Zhao study[9] had no correlation with ligand efficiency ($\tau = -0.02$, $p \approx 0.15$).

The OPLS4 per-atom strain estimates for the 10% PDBBind subset maintained a significant inverse correlation with ligand efficiency ($\tau = -0.29$, $p \ll 0.001$, 95% CI: −0.37 to −0.22), though reduced slightly in magnitude from the correlations seen with the xGen strain estimates. It is likely that a fully native real-space ligand refinement protocol that employed OPLS4 would reduce strain estimates from the *post hoc* calculation done here. It would be interesting to see whether the correlation to LEFF would increase in magnitude, with the expectation that employing more extensive and nominally accurate torsional parameterization during ligand refinement should lead to a better correlation with experimental measurements.

Ligand Efficiency, Strain, and Design Strategy. Table 1 lists the examples from Figure 3 in order of decreasing ligand efficiency. For context, the median LEFF for the full PDBBind Set was 0.35, with the 95th and 5th percentiles being 0.80 and 0.18, respectively. Four cases had high binding site contact, and two had low (highlighted with italicized contact and upper bound values). The biotin/streptavidin interaction is well-known as an extreme example of ligand efficiency, even with the mutant S45A protein variant (Panel A of Figure 3), achieving LEFF of 0.83. With the wild-type protein, LEFF exceeds 1.0 kcal/mol/atom.[32] Estimated strain was 0.0 kcal/mol, as is common among the most highly efficient ligands.

The next two cases fall into a common category, with over 20% of observed LEFF being in the range of 0.25−0.40 kcal/mol/atom. The DDR1 kinase inhibitor (6GWR, Panel B of Figure 3) has low strain and high binding site contact, but only moderate efficiency. Improvement strategies would likely need to focus on the direct binding interaction. The ligand of 1PZI (Panel C), by contrast, could likely be improved by engineering a lower strain analog, but it also suffers in terms of efficiency by being at the border of high and low binding site contact.

The fourth case (Panel D of Figure 3) is a macrocyclic ligand whose alkyl linker is solvent-exposed (exhibiting conformational variation in the xGen ensemble) and whose binding site contact is low. Strain was also very low from the xGen ensemble, suggesting that modifications to the protein−ligand interaction footprint might be a better strategy than strain reduction. In this case, rigidification of the macrocyclic linker might also be

https://doi.org/10.1021/acs.jmedchem.2c01744
J. Med. Chem. 2023, 66, 1955−1971

Table 1. Example Ligands from Figure 3

PDB Code	Ligand	Protein	*N* Non-H	Contact	Δ*G*	xGen Strain	Model UB_{95}	LEFF	Strain/*N*
1DF8	Biotin	Streptavidin S45A	16	1.00	−13.2	0.0	2.2	0.83	0.00
6GWR	Synthetic	DDR1 Kinase	41	0.92	−11.0	1.1	9.8	0.27	0.03
1PZI	Synthetic	Enterotoxin B	23	0.75	−5.7	2.3	3.9	0.25	0.10
5T9W	Synthetic	PPIA Cyclophilin	40	*0.67*	−7.6	1.6	5.9	0.19	0.04
3EOV	Cyclosporin	*L. donovani* Cyclophilin	85	0.41	−10.6	6.3	*17.2*	0.12	0.07
5GLH	Endothelin	ETB Receptor	171	0.74	−14.6	28.3	55.4	0.09	0.17

valuable. Note that the estimated strain from the deposited PDB coordinates was 7.1 kcal/mol, which might lead to a different conclusion for optimization, as would the static picture of the alkyl linker.

The final two examples are both naturally occurring peptidic macrocycles. Both represent extrema, with cyclosporin exhibiting very low binding site contact and endothelin having high binding site contact for an extremely large ligand. Both exhibit poor LEFF (below the 5th percentile) but for different reasons: very low contact for the former and moderately high strain for the latter. The case of cyclosporin is unusual across many different binding interactions, forming a vertical column of low-strain data points (at atom count 85) in Figures 5 and 9, distorting the sample strain estimates somewhat. Perhaps little can be learned from the endothelin/endothelin-receptor interaction from a design perspective, as that system is likely to have co-evolved.

The superlinear dependence of ligand strain on ligand size does not completely explain the inverse correlation between LEFF and ligand size. Figure 12 shows the relationship of LEFF

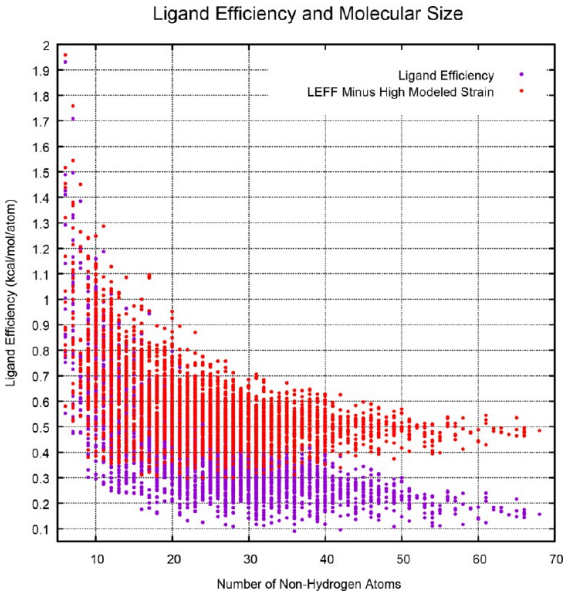


Figure 12. Relationship of LEFF to ligand size (violet) and LEFF adjusted upward to account for the modeled high-strain upper bound at each molecular size (red).

to ligand size (violet points, as in Figure 11) along with an adjusted plot of LEFF (red points). The adjustment simulated eliminating the near-maximal effect of ligand strain on LEFF. In the adjusted plot, LEFF values were shifted assuming that each ligand incurred the upper bound of high-contact modeled strain, with the strain value per-atom added to each data point. The inverse correlation between ligand size and LEFF decreased in magnitude to $\tau = -0.35$ ($p \ll 0.001$, 95% CI: −0.38 to −0.33) from $\tau = -0.55$ ($p \ll 0.001$, 95% CI: −0.57 to −0.54). Design strategies to drastically reduce ligand strain should not be expected to eliminate low LEFF in large ligands, even in the most optimistic strain reduction scenario, particularly in cases of low binding site contact.

CONCLUSIONS

Estimating bound ligand strain cannot be done meaningfully if the modeled coordinates of the ligand bound to the target of interest nominally fit electron density but much lower energy ligand models exist that fit equally well. No level of energetic theory can overcome poorly fit ligands. There have been a number of methods introduced to improve ligand fitting in X-ray diffraction data,[14−19] and some have been shown to improve ligand energetics. The xGen method used here has been validated on a broad variety of ligands, particularly complex macrocycles,[20,21] and in the present work an additional set of nearly 3,000 protein−ligand complexes. Associated strain estimates are consistently lower (roughly 40%) than those obtained using positionally restrained, classically modeled ligand coordinates. Further, they are quantitatively better correlated with ligand efficiency.

We have derived a distributional model for ligand strain that fits empirical data remarkably well when accounting for the degree of ligand binding site contact, with the vast majority of ligands falling to the high-contact category of binding. The predicted upper bound on ligand strain offers a tighter constraint for conformational search protocols than is typically used. For example, ligands with up to 25 non-hydrogen atoms have a well-modeled 95th percentile upper bound of just 4.5 kcal/mol, and the upper bound for 40 atoms is 9.4 kcal/mol. Importantly, the model is not dependent on molecular class, whether "typical" small molecules, non-peptidic macrocycles, or peptidic macrocycles. It is only dependent on molecular size. It is not uncommon for computational chemists supporting lead identification or optimization efforts to generate ligand conformation ensembles spanning energy windows of 15−20 kcal/mol above the identified force field minimum. For many molecules in typical size ranges, windows of this size unnecessarily expand conformer pools to the detriment of denser sampling of lower energy conformers.

Apart from the practical utility of an accurate upper bound on ligand strain, the existence of a distributional model has implications for design strategy. If, for example, a lead project compound has relatively high strain compared with expectations, strategies to reduce conformational strain while maintaining the existing interaction footprint are likely to succeed, resulting in improved potency. Conversely, if estimated strain is already at the lower end of the distribution for the lead compound's size, efforts to engineer lower strain would probably be better spent toward modifying the protein−ligand interaction directly.

https://doi.org/10.1021/acs.jmedchem.2c01744
J. Med. Chem. 2023, 66, 1955−1971

EXPERIMENTAL SECTION

Molecular Data Sets. The Brueckner Set contained 341 protein−ligand complexes: 38 non-macrocycles from the Perola/Charifson study,[3] 147 non-peptidic macrocycles,[20] and 156 peptidic macrocycles.[21] Description of the curation and characteristics of those sets has been previously detailed.

The PDBBind 2020 Refined Set contains 5,316 entries. Automatic processing of the PDB files to separate protein, water, and ligands along with assigning bond orders, adding protons, and checking ligand connectivity against SMILES strings associated with the annotated ligand HET codes was done using the *getpdb* and *grindpdb* Surflex Platform commands. Of those that passed the initial automatic processing, 4,052 had electron density available. Complexes where the PDB deposited ligand coordinates were clearly poor (RSCC < 0.65 or RSR > 0.45) were eliminated, resulting in 3,942 complexes. Real-space refinement was carried out using the xGen "diverse ensemble" protocol,[20] which produces the highest nominal ligand ensemble fit quality but allows for quite large conformational ensembles. Complexes were retained where the diverse xGen ligand conformer ensemble fits were high (RSCC ≥ 0.85 and RSR ≤ 0.25), resulting in 2,996 final PDBBind Set complexes.

In order to avoid conformers having extremely low occupancy, the PDBBind Set of 2,996 complexes was again refined with the xGen "strict ensemble" protocol, which produced 2.5 conformers on average. Overall, the real-space fit quality was high for both the xGen ensembles and the original PDB ligand models. Improvements in RSCC and RSR using the xGen ensemble approach were small but consistent compared with original PDB ligand coordinates: an average RSCC of 0.924 and an average RSR of 0.156, compared to the original PDB ligand coordinate metrics of 0.897 and 0.169, respectively. With the xGen approach, 95% of ligands had RSCC ≥ 0.87 and RSR ≤ 0.22. Respectively, the original PDB ligand models had RSCC ≥ 0.82 and RSR ≤ 0.24. The same protocol was run for the Brueckner Set, with similarly consistent improvements in RSCC and RSR.[20,21]

Covalent ligands have been excluded in the curation of the data sets, because they present numerous difficulties for both ligand refinement and interpretation of binding affinity data. Binding sites with tightly held metal ions have not been excluded, due to their presence in many pharmaceutically relevant targets. These cases might lead to higher ligand strain due to the strength of chelation interactions, but they comprise less than 15% of the cases overall, and they represent examples of non-covalent interactions that are genuinely important.

Molecular conformational complexity was calculated based on the output of the *regen3d* command within the Tools Module of the Surflex Platform. The output includes information about the number of non-hydrogen atoms and counts for exocyclic and macrocyclic rotatable bonds.

Note that all compounds described in this study have been previously disclosed and have been characterized by X-ray crystallographic experiment, with details regarding purity being available in the original publications or associated with the RCSB PDB deposition codes.

Conformer Neighborhoods and Energy-Surrogates. A key aspect of ensemble generation is the idea of conformer neighborhoods as shown in Figures 1 and 4. Ideal fit to electron density generally yields distortions from ideal ligand geometry, so the initial refinement procedure balances the two considerations, achieving reasonable density fit while avoiding large geometric distortions. As described earlier, each member of this balanced-pool conformer set is minimized in two ways: a) under a condition in which the density overlap is strongly weighted and b) with no density overlap reward but with a positional restraint (the square-welled quadratic penalty is 1.0 kcal/mol/Å^2 for deviations beyond 0.2 Å). These three pools are used to identify conformer trios with the property that a low-energy "min-pool" member and a high-fit-quality "density-pool" member are both within a small neighborhood of a central "balanced-pool" member.

One could use conventional RMSD to define the extents of such a neighborhood, but this becomes problematic across different molecular sizes. An RMSD of 0.25 Å for two conformers of a molecule with 6 heavy atoms yields a maximal single atomic deviation of 0.61 Å. But for a molecule with 60 heavy atoms, the same 0.25 Å RMSD can mean a single atomic deviation of as much as 1.9 Å. As previously described,[20] we define the sRMSD (scaled RMSD) as the average of the conventional RMSD and the single maximum atomic deviation. This sRMSD is always at least as large as the conventional RMSD. We use 0.65 Å sRMSD as the threshold neighborhood size (the maximum sRMSD from the central member to any other member). The sRMSD of 0.65 Å generally corresponds to a conventional RMSD of 0.3−0.4 between members of a single neighborhood. Figure 13 shows the full neighborhood around a single balanced-pool conformer for the 3SUE ligand, which contains 20 from the min-pool (salmon) and 34 from the density-pool (orange).

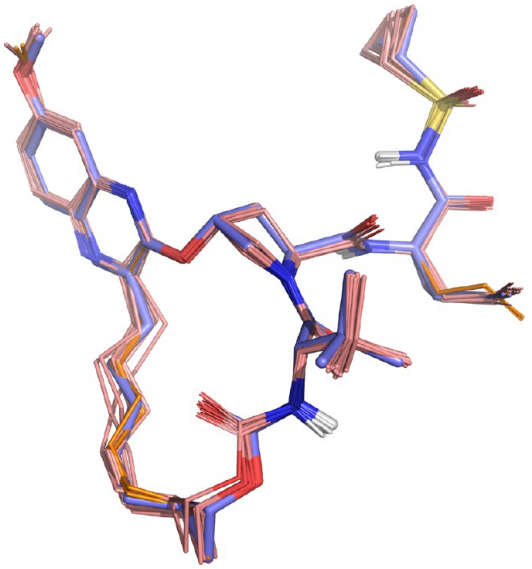


Figure 13. A full conformer neighborhood: a single balanced-pool conformer (slate), the full set of density-pool (orange) and min-pool conformers (salmon) within 0.65 Å sRMSD of the balanced conformer.

One could use the min-pool conformers corresponding to those that form xGen ensemble (which come from the density-pool) directly as energy-surrogates and report the minimum energy (or a weighted energy) as the bound conformational strain. However, we perform an additional positionally restrained minimization for two reasons.

First, nominal local minima can be frustrated by the bumpiness of force field energy surfaces. This can be overcome by performing multiple, slightly perturbed, positionally restrained minimizations. Here, we have used 5 small perturbations of 0.01 Å for each atom along with a square-welled quadratic penalty of 1.0 kcal/mol/Å^2 for deviations beyond 0.1 Å from the original conformer's atomic position. This typically results in small improvements in energy (approximately 0.3 kcal/mol) but 2% exhibit improvements of 2.0 kcal/mol or greater. The resulting final conformers remain very close to the original crystallographically optimized xGen conformers (mean RMSD of 0.3 Å).

Second, when employing an alternative force field, this final positionally restrained minimization offers a straightforward means to make direct comparisons. We employed an analogous procedure for OPLS4[31] on a randomly selected 10% of the PDBBind Data Set (299 cases). Positionally restrained minimizations were carried out on the strict xGen conformer ensembles *and* their corresponding min-pool conformer-trio siblings, with the lowest energy being recorded as that for the bound ligand state. Minimizations were carried out with a square-welled quadratic positional restraint on non-hydrogen atoms (1.0 kcal/mol/Å^2 for deviations beyond 0.3 Å). This was a slightly looser restraint than used for the xGen protocol, owing to differences in termination criteria in the optimization algorithms. RMS deviations for the OPLS4 optimized conformers from the fitted ensemble conformers

https://doi.org/10.1021/acs.jmedchem.2c01744
J. Med. Chem. 2023, 66, 1955−1971

averaged 0.3 Å, consistent with the xGen strain estimation protocol. Minimizations were performed in the gas phase, with a dielectric of 80.0, in order to parallel the xGen protocol.

For all xGen real-space refinement calculations and related strain estimates, we employed version 5.1 of the Surflex Platform (BioPharmics LLC, Sonoma County, CA). For all OPLS4 calculations, we employed Batchmin MacroModel v13.6 of the Schrödinger 2022-2 Suite (Schrödinger, LLC, New York, NY, 2019).

Global Minimum Determination. Global strain is calculated based on the difference between the bound-state energy (from the energy-surrogate calculation above) and the unbound-state minimum energy, which is the global minimum energy from an exhaustive conformational search of the ligand. This is calculated using the ForceGen conformational search method, which has been previously described[34,35]

For small, drug-like molecules, the *-pquant* level of conformational elaboration is likely to be sufficient to identify global minima in the vast majority of cases, based on the greater than 90% success rate of identifying close-to-crystallographic conformers (≤1.0 Å RMSD) beginning from random starting conformations. However, particularly for large, peptidic macrocycles, we adopted an iterative approach to conformational search in order to better ensure adequate sampling.[21] This iterative search has been implemented as a command within the Tools Module of the Surflex Platform, called *fgen_deep*.

Beginning from a single input conformer, the *fgen_deep* procedure performs a standard ForceGen search, with the resulting conformer pool being clustered by RMSD. If the resulting *N* lowest-energy clusters contain new conformations compared with prior rounds, search is iterated beginning with the lowest energy conformers from each of the *N* new clusters. Multiple rounds of this are carried out, each time consolidating the full set of conformers into a non-redundant set within a specified energetic window prior to clustering. The process is iterated until no new low-energy clusters are identified.

Here, for all but the peptidic macrocycles (which had been subjected to iterative search previously), the *fgen_deep* method was employed to identify global minima. The lowest-energy 100 conformers from each respective pool were subjected to unconstrained tweaked minimization to identify the final estimate of the global minimum. For the PDBBind Set, the average number of conformers in the final *fgen_deep* pools was 6,953 (the range was 1−30,866).

For the OPLS4 comparison, the lowest-energy 1,000 conformers from the *fgen_deep* procedure were subjected to unconstrained minimization, with the lowest energy being recorded as the global minimum. Minimizations were performed in the gas phase, with a dielectric of 80.0, in order to parallel the xGen protocol.

Ligand Efficiency Definition. Ligand efficiency is an old concept, dating back at least to Kuntz and Kollman's discussions about the maximal affinity of ligands.[32] Some controversy has arisen over time regarding definitions. We define ligand efficiency as follows:

$$\Delta G^{\circ} = -RT \ln\left(\frac{K_d}{C^{\circ}}\right) \tag{6}$$

$$\mathrm{LE} = -\frac{\Delta G^{\circ}}{N_{nonH}} = -\frac{1.36\ \mathrm{p}K_d}{N_{nonH}} \tag{7}$$

The above definition follows the treatment of Hopkins et al.,[36] where standard choices for temperature, pressure, and concentration in thermodynamic systems are 298 K, 1 atm, and 1 M. Common practice equates pK_d with pK_i, pIC_{50}, and pEC_{50}, allowing for calculation of LE for all of the PDBBind data, with the normal caveats about comparisons of IC_{50} values across different assay conditions.

Note that Kenny[37] has highlighted the arbitrary choice of the standard concentration C° to be 1 M and that changing that value has complex effects on the interpretation of ligand efficiency. However, the choice is uncontroversial, and the definition in eq 7 is widely accepted.

■ ASSOCIATED CONTENT

Data Availability Statement

The Supporting Information includes a detailed description of the contents of the extensive data archive (freely downloadable at www.jainlab.org). The archive contains a summary spreadsheet of calculated strain values along with data for all protein−ligand complexes, including protein structures, original PDB ligand models, xGen fitted conformer ensembles (and associated conformer trio members), and scripts to calculate estimated strain energy values from the provided data. All software employed herein is commercially available.

Supporting Information

The Supporting Information is available free of charge at https://pubs.acs.org/doi/10.1021/acs.jmedchem.2c01744.

Descriptions of the molecular data sets, data archive, and linear distributional model (PDF)

Jain-Sherer-Strain-SI file, containing tabular summary information for all of the complexes (XLSX)

■ AUTHOR INFORMATION

Corresponding Authors

Ajay N. Jain − *Research & Development, BioPharmics LLC, Sonoma County, California 95404, United States;* orcid.org/0000-0003-4641-8501; Email: ajain@jainlab.org

Edward C. Sherer − *Analytical Research and Development, Merck & Co. Inc., Rahway, New Jersey 07065, United States;* orcid.org/0000-0001-8178-9186; Email: edward_sherer@merck.com

Authors

Alexander C. Brueckner − *Molecular Structure & Design, Bristol Myers Squibb, Princeton, New Jersey 08543, United States*

Ann E. Cleves − *Research & Development, BioPharmics LLC, Sonoma County, California 95404, United States;* orcid.org/0000-0002-1622-2770

Mikhail Reibarkh − *Analytical Research and Development, Merck & Co. Inc., Rahway, New Jersey 07065, United States;* orcid.org/0000-0002-6589-707X

Complete contact information is available at: https://pubs.acs.org/10.1021/acs.jmedchem.2c01744

Notes

The authors declare no competing financial interest.

■ ACKNOWLEDGMENTS

The authors thank Juan Alvarez, Qiaolin Deng, Qi Gao, Chip Lesburg, and Steve Soisson for helpful discussions.

■ ABBREVIATIONS USED

RSCC, real-space correlation coefficient; RSR, real-space R; LEFF, ligand efficiency; VdW, van der Waals

■ REFERENCES

(1) Nicklaus, M. C.; Wang, S.; Driscoll, J. S.; Milne, G. W. Conformational Changes of Small Molecules Binding to Proteins. *Bioorg. Med. Chem.* **1995**, *3*, 411−428.

(2) Boström, J.; Norrby, P.-O.; Liljefors, T. Conformational Energy Penalties of Protein-Bound Ligands. *J. Comput.-Aided Mol. Des.* **1998**, *12*, 383−383.

(3) Perola, E.; Charifson, P. S. Conformational analysis of drug-like molecules bound to proteins: An extensive study of ligand reorganization upon binding. *J. Med. Chem.* **2004**, *47*, 2499−2510.
(4) Butler, K. T.; Luque, F. J.; Barril, X. Toward accurate relative energy predictions of the bioactive conformation of drugs. *J. Comput. Chem.* **2009**, *30*, 601−610.
(5) Foloppe, N.; Chen, I.-J. Conformational sampling and energetics of drug-like molecules. *Curr. Med. Chem.* **2009**, *16*, 3381−3413.
(6) Fu, Z.; Li, X.; Merz, K. M., Jr Accurate assessment of the strain energy in a protein-bound drug using QM/MM X-ray refinement and converged quantum chemistry. *J. Comput. Chem.* **2011**, *32*, 2587−2597.
(7) Sitzmann, M.; Weidlich, I. E.; Filippov, I. V.; Liao, C.; Peach, M. L.; Ihlenfeldt, W.-D.; Karki, R. G.; Borodina, Y. V.; Cachau, R. E.; Nicklaus, M. C. PDB Ligand Conformational Energies Calculated Quantum-Mechanically. *J. Chem. Inf. Model.* **2012**, *52*, 739−756.
(8) Avgy-David, H. H.; Senderowitz, H. Toward focusing conformational ensembles on bioactive conformations: A molecular mechanics/quantum mechanics study. *J. Chem. Inf. Model.* **2015**, *55*, 2154−2167.
(9) Tong, J.; Zhao, S. Large-Scale Analysis of Bioactive Ligand Conformational Strain Energy By Ab Initio Calculation. *J. Chem. Inf. Model.* **2021**, *61*, 1180−1192.
(10) Zivanovic, S.; Colizzi, F.; Moreno, D.; Hospital, A.; Soliva, R.; Orozco, M. Exploring the Conformational Landscape of Bioactive Small Molecules. *J. Chem. Theory Comput.* **2020**, *16*, 6575−6585.
(11) Liebeschuetz, J.; Hennemann, J.; Olsson, T.; Groom, C. R. The good, the bad and the twisted: A survey of ligand geometry in protein crystal structures. *J. Comput.-Aided Mol. Des.* **2012**, *26*, 169−183.
(12) Liebeschuetz, J. W. The Good, the Bad, and the Twisted Revisited: An Analysis of Ligand Geometry in Highly Resolved Protein-Ligand X-ray Structures. *J. Med. Chem.* **2021**, *64*, 7533−7543.
(13) Reynolds, C. H. Protein−ligand cocrystal structures: We can do better. *ACS Med. Chem. Lett.* **2014**, *5*, 727−729.
(14) Wlodek, S.; Skillman, A.; Nicholls, A. Automated ligand placement and refinement with a combined force field and shape potential. *Acta Crystallogr., Sect. D: Biol. Crystallogr.* **2006**, *62*, 741−749.
(15) Janowski, P. A.; Moriarty, N. W.; Kelley, B. P.; Case, D. A.; York, D. M.; Adams, P. D.; Warren, G. L. Improved ligand geometries in crystallographic refinement using AFITT in PHENIX. *Acta Crystallogr., Sect. D: Struct. Biol.* **2016**, *72*, 1062−1072.
(16) Van Zundert, G. C.; Hudson, B. M.; de Oliveira, S. H.; Keedy, D. A.; Fonseca, R.; Heliou, A.; Suresh, P.; Borrelli, K.; Day, T.; Fraser, J. S.; van den Bedem, H. qFit-ligand reveals widespread conformational heterogeneity of drug-like molecules in X-ray electron density maps. *J. Med. Chem.* **2018**, *61*, 11183−11198.
(17) Riley, B. T.; Wankowicz, S. A.; de Oliveira, S. H.; van Zundert, G. C.; Hogan, D. W.; Fraser, J. S.; Keedy, D. A.; van den Bedem, H. qFit 3: Protein and Ligand Multiconformer Modeling for X-ray Crystallographic and Single-Particle Cryo-EM Density Maps. *Protein Sci.* **2021**, *30*, 270−285.
(18) Borbulevych, O. Y.; Plumley, J. A.; Martin, R. I.; Merz, K. M.; Westerhoff, L. M. Accurate macromolecular crystallographic refinement: Incorporation of the linear scaling, semiempirical quantum-mechanics program DivCon into the PHENIX refinement package. *Acta Crystallogr., Sect. D: Biol. Crystallogr.* **2014**, *70*, 1233−1247.
(19) Borbulevych, O.; Martin, R. I.; Westerhoff, L. M. High-throughput quantum-mechanics/molecular-mechanics (ONIOM) macromolecular crystallographic refinement with PHENIX/DivCon: the impact of mixed Hamiltonian methods on ligand and protein structure. *Acta Crystallogr., Sect. D: Struct. Biol.* **2018**, *74*, 1063−1077.
(20) Jain, A. N.; Cleves, A. E.; Brueckner, A. C.; Lesburg, C. A.; Deng, Q.; Sherer, E. C.; Reibarkh, M. Y. XGen: Real-Space Fitting of Complex Ligand Conformational Ensembles to X-Ray Electron Density Maps. *J. Med. Chem.* **2020**, *63*, 10509−10528.
(21) Brueckner, A. C.; Deng, Q.; Cleves, A. E.; Lesburg, C. A.; Alvarez, J. C.; Reibarkh, M. Y.; Sherer, E. C.; Jain, A. N. Conformational Strain of Macrocyclic Peptides in Ligand−Receptor Complexes Based on Advanced Refinement of Bound-State Conformers. *J. Med. Chem.* **2021**, *64*, 3282−3298.
(22) Tickle, I. J. Statistical Quality Indicators for Electron-Density Maps. *Acta Crystallogr., Sect. D: Biol. Crystallogr.* **2012**, *68*, 454−467.
(23) Romano, K. P.; Ali, A.; Aydin, C.; Soumana, D.; Özen, A.; Deveau, L. M.; Silver, C.; Cao, H.; Newton, A.; Petropoulos, C. J.; et al. The Molecular Basis of Drug Resistance Against Hepatitis C Virus NS3/4A Protease Inhibitors. *PLoS Pathogens* **2012**, *8*, e1002832.
(24) Liu, Z.; Li, Y.; Han, L.; Li, J.; Liu, J.; Zhao, Z.; Nie, W.; Liu, Y.; Wang, R. PDB-wide collection of binding data: Current status of the PDBbind database. *Bioinformatics* **2015**, *31* (3), 405−412.
(25) Liu, Z.; Su, M.; Han, L.; Liu, J.; Yang, Q.; Li, Y.; Wang, R. Forging the Basis for Developing Protein−Ligand Interaction Scoring Functions. *Acc. Chem. Res.* **2017**, *50*, 302−309.
(26) Venugopal, V.; Datta, A. K.; Bhattacharyya, D.; Dasgupta, D.; Banerjee, R. Structure of Cyclophilin from Leishmania donovani Bound to Cyclosporin at 2.6 Å Resolution: Correlation Between Structure and Thermodynamic Data. *Acta Crystallogr., Sect. D: Biol. Crystallogr.* **2009**, *65*, 1187−1195.
(27) Shihoya, W.; Nishizawa, T.; Okuta, A.; Tani, K.; Dohmae, N.; Fujiyoshi, Y.; Nureki, O.; et al. Activation Mechanism of Endothelin ETB Receptor by Endothelin-1. *Nature* **2016**, *537*, 363−368.
(28) Foloppe, N.; Chen, I.-J. Towards understanding the unbound state of drug compounds: Implications for the intramolecular reorganization energy upon binding. *Bioorg. Med. Chem.* **2016**, *24*, 2159−2189.
(29) Reynolds, C. H.; Tounge, B. A.; Bembenek, S. D. Ligand binding efficiency: Trends, physical basis, and implications. *J. Med. Chem.* **2008**, *51*, 2432−2438.
(30) Meyder, A.; Nittinger, E.; Lange, G.; Klein, R.; Rarey, M. Estimating Electron Density Support for Individual Atoms and Molecular Fragments in X-ray Structures. *J. Chem. Inf. Model.* **2017**, *57*, 2437−2447.
(31) Lu, C.; Wu, C.; Ghoreishi, D.; Chen, W.; Wang, L.; Damm, W.; Ross, G. A.; Dahlgren, M. K.; Russell, E.; Von Bargen, C. D.; Abel, R.; Friesner, R. A.; Harder, E. D. OPLS4: Improving force field accuracy on challenging regimes of chemical space. *J. Chem. Theory Comput.* **2021**, *17*, 4291−4300.
(32) Kuntz, I.; Chen, K.; Sharp, K.; Kollman, P. The Maximal Affinity of Ligands. *Proc. Natl. Acad. Sci. U. S. A.* **1999**, *96*, 9997−10002.
(33) Reynolds, C. H.; Reynolds, R. C. Group additivity in ligand binding affinity: An alternative approach to ligand efficiency. *J. Chem. Inf. Model.* **2017**, *57*, 3086−3093.
(34) Cleves, A. E.; Jain, A. N. ForceGen 3D Structure and Conformer Generation: From Small Lead-Like Molecules to Macrocyclic Drugs. *J. Comput.-Aided Mol. Des.* **2017**, *31*, 419−439.
(35) Jain, A. N.; Cleves, A. E.; Gao, Q.; Wang, X.; Liu, Y.; Sherer, E. C.; Reibarkh, M. Y. Complex macrocycle exploration: Parallel, heuristic, and constraint-based conformer generation using ForceGen. *J. Comput.-Aided Mol. Des.* **2019**, *33*, 531−558.
(36) Hopkins, A. L.; Keserü, G. M.; Leeson, P. D.; Rees, D. C.; Reynolds, C. H. The Role of Ligand Efficiency Metrics in Drug Discovery. *Nat. Rev. Drug Discovery* **2014**, *13*, 105−121.
(37) Kenny, P. W. The Nature of Ligand Efficiency. *J. Cheminform.* **2019**, *11*, 9.

https://doi.org/10.1021/acs.jmedchem.2c01744
J. Med. Chem. 2023, 66, 1955−1971