JACS
JOURNAL OF THE AMERICAN CHEMICAL SOCIETY
Article
Cite This: J. Am. Chem. Soc. 2019, 141, 17142−17149
pubs.acs.org/JACS

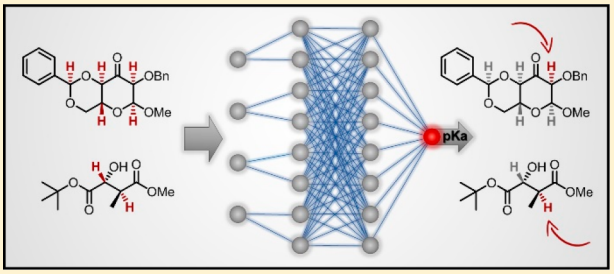
Rapid and Accurate Prediction of pK_a Values of C−H Acids Using Graph Convolutional Neural Networks

Rafał Roszak,[†,§,⊥] Wiktor Beker,[†,§,⊥] Karol Molga,[†] and Bartosz A. Grzybowski*[,†,‡,§]

[†]Institute of Organic Chemistry, Polish Academy of Sciences, ul. Kasprzaka 44/52, 01-224 Warsaw, Poland
[‡]Institute for Basic Science, Center for Soft and Living Matter, Ulsan 44919, South Korea
[§]Allchemy, Inc., 2145 45th Street #201, Highland, Indiana 46322, United States

Supporting Information

ABSTRACT: The ability to estimate the acidity of C−H groups within organic molecules in non-aqueous solvents is important in synthetic planning to correctly predict which protons will be abstracted in reactions such as alkylations, Michael additions, or aldol condensations. This Article describes the use of the so-called graph convolutional neural networks (GCNNs) to perform such predictions on the time scales of milliseconds and with accuracy comparing favorably with state-of-the-art solutions, including commercial ones. The crux of the method is to train GCNNs using descriptors that reflect not only topological but also chemical properties of atomic environments. The model is validated against adversarial controls, supplemented by the discussion of realistic synthetic problems (on which it correctly predicts the most acidic protons in >90% of cases), and accompanied by a Web application intended to aid the community in everyday synthetic planning.



1. INTRODUCTION

Since pK_a values dictate the ionization states of both proteins[1] and drug molecules,[2a−d] most effort in predicting these values has so far focused on groups that are ionizable under physiological conditions (e.g., NH or OH) for which large quantities of data are available. This effort has, indeed, met with considerable success, and state-of-the-art solutions such as Bayer's "S+pKa" model[2a]—developed using data for >25 000 compounds and based on neural network ensembles and analysis of microstates—produce mean average errors well below 1 pK_a unit. The problem is, however, much wider, and pK_a prediction is also very important in organic synthesis to foretell outcomes of reactions involving proton abstraction and formation of carbanion intermediates (e.g., in alkylations, Michael additions, or aldol condensations). Examples of failed reactions or "unexpected" reaction outcomes described in the synthetic literature[3a−c] (Figure 1) indicate that prediction of pK_a values of C−H acids in non-aqueous solvents is not straightforward and often cannot be addressed based on one's chemical intuition and a handful of values tabulated[4] for common functional groups.

Although quantum-mechanical (QM) models can offer accurate pK_a estimates (with mean absolute error (MAE) of 1.3 pK_a units when combined with empirical correction[5]), these methods entail considerable computational times (e.g., for Schrödinger's commercial Jaguar software[6] run on a multicore desktop, minutes to hours depending on molecule size) and the accuracy-improving empirical corrections are available only for series of structurally related compounds (for more detailed discussion of existing models, both *ab initio*,[7a−g] and statistical,[8a−e] see Supporting Information (SI), Section S1), making such solutions rarely used in synthetic practice. In addition, the time scales of QM-based calculations are incompatible with high-throughput applications including the emerging retrosynthesis software platforms (our Chematica,[9a] MIT's ASKCOS,[9b] Waller's MCTS[9c]) which inspect thousands of reaction candidates per minute to construct viable synthesis plans. In such programs, there is currently no means of rapidly determining the most acidic protons and, based on our experience, a large proportion of erroneous synthetic plans reflect lack of proper pK_a treatment. Unfortunately, no methods for rapid yet reasonably accurate pK_a prediction in non-aqueous solvents are currently available; to our knowledge, the state-of-the-art is Schrödinger's commercial Epik package,[10a,b] for which, however, the MAE, quantified against experimental values is as large as 3.3 pK_a units (Figure 2; red markers).

Motivated by these considerations and current limitations, we use here the so-called graph convolutional neural networks, GCNNs, which have been previously implemented[11] to predict molecule-wide properties and outcomes/loci of organic reactions but not atomic properties such as pK_a. The GCNNs we implement capture not only topological but also chemical neighborhoods of individual atoms within a molecule, allowing us to train a model that estimates pK_a(DMSO) values of C−H acids in arbitrary organic molecules within milliseconds, with MAE ≈ 2.1 pK_a units (Figure 2; green markers),

Received: June 6, 2019
Published: October 21, 2019

DOI: 10.1021/jacs.9b05895
J. Am. Chem. Soc. 2019, 141, 17142−17149

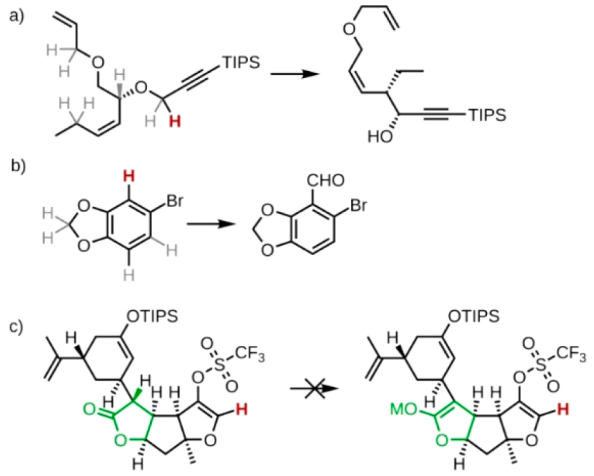


Figure 1. Examples of reactions whose non-obvious outcomes are dictated by acidity of C−H protons. (a) Regioselectivity of 2,3-Wittig rearrangement depends on the relative acidity of allylic/propargylic protons. In this example, three deprotonation sites are, in principle, possible (protons colored gray) but reaction occurs selectively at the most acidic, lowest-pK_a proton colored in red.[3a] (b) 1,2-Methylenedioxy is an ortho-directing metalation group. However, the presence of bromine in the aromatic ring makes pK_a of both ortho positions different enough for selective lithiation. In experiment, treatment with LDA and quenching with DMF gave only the regioisomer of benzaldehyde shown in the figure[3b] (and involving proton colored red). (c) At the late stage of total synthesis of Ineleganolide, deprotonation and epimerization at position α to the carbonyl group colored green was attempted. Unfortunately, the desired product was not observed due to higher acidity of the vinylic proton colored in red. Because of this unforeseen acidity, the eight-year-long project failed.[3c]

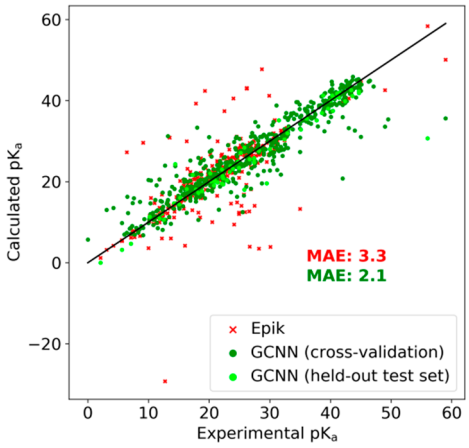


Figure 2. Comparison of pK_a predictions on Schrödinger's Epik (red markers) and our GCNN model (green markers) against a set of experimentally measured values.[4] Dark-green markers correspond to the GCNN results from the standard 5-fold cross-validation of the 90% of the dataset (i.e., this subset was randomly divided into five parts and predictions for each part were made on the model trained on the remaining four parts). Light-green markers are for the remaining 10%, held-out test set.

and with >90% correct predictions of the outcomes of ∼13 000 diverse reactions involving proton abstraction. Interestingly, the model performs well only if the descriptors used to train it reflect physical quantities such as Pauling's electronegativity or Gasteiger's partial charges—this result echoes our previous findings (in the context of predicting the outcomes of pericyclic reactions[12]) that development of accurate and transferable (beyond similar compounds) AI models requires the use of chemically meaningful features. Furthermore, the GCNN model is supplemented by routines addressing cases when deprotonation is controlled by the equivalents of base used, pre-coordination of the base, or the strain in the substrate. All these features are incorporated into a Web application freely available at https://pka.allchemy.net which, we hope, will become a useful resource aiding organic chemists in everyday synthesis design.

2. RESULTS AND DISCUSSION

2.1. Selection of the Training Set. Our solvent of choice was DMSO because of the largest number (compared to other solvents) of available experimental pK_a values,[4] and because pK_a measured in that solvent can be rescaled to other polar aprotic organic solvents, like THF or acetonitrile (e.g., $pK_a^{THF} = -0.963 + 1.046pK_a^{DMSO}$, ref 13; $pK_a^{MeCN} = 11.6 + 0.98pK_a^{DMSO}$, ref 8e). Still, the number of such experimental values is relatively small (414) and so we sought to extend this set—to be later used for training of our neural networks—via the linear free energy relationship (LFER), which assumes linear dependence between pK_a and calculated dissociation free energy, $pK_a = a\Delta G + b$, with empirical constants a and b obtained from linear regression to experimental data.[7f] Within the series of homologous compounds, the zero-point energy and entropy differences between neutral and anionic species are virtually constant.[7g] Hence, the free energy of proton dissociation can be approximated as $\Delta G \approx \Delta G_{solv} + \Delta E_{gas}$ and the resulting small systematic error is corrected by constant b. Importantly, if properly tuned,[7h] LFER allows for establishing accurate pK_a values within series of analogous compounds. In our 414 dataset, nitroalkanes (R-NO_2), nitriles (R-CN), alkyl esters (R-COOR), and sulfones (R-SO_2Ph) had enough experimental data points, differing in only one substituent, to establish the a and b parameters for each class separately (14, 15, 11, and 26 examples), allowing for the use of LFER approach in a meaningful manner.

Specifically, for each molecule from these classes, 1000 conformers for both neutral and anionic species were generated, their energies were approximated roughly by molecular mechanics (MMFF[14a] and UFF[14b]), and 100 lowest-energy conformers were chosen for detailed QM calculations performed under DFT level of theory with polarizable continuum model (PCM) to include solvation effects. DFT functional, basis set, and parameters of the molecular cavity used in the implicit solvent calculations were optimized toward the best correlation with the experimental pK_a values of the largest and thus the most significant sulfone series. After examination of about 50 functionals, 26 basis sets, and 22 implicit solvent parameters, we ultimately selected HISSbPBE/def2tzvp with molecular cavity defined by atomic UFF radii scaled by the factor of 1.5 (for further details, see SI, Section S2). The energies corresponding to the optimized geometries were used to fit the LFER trends (solid markers in Figure 3), in each case with correlation coefficients $R^2 > 0.92$. Subsequently, similar QM calculations were performed to determine pK_a's of 212 manually curated, homologous molecules which were added into and fitted to these series. Overall, these calculations took several months of CPU time on the TASK and ICM supercomputers located in, respectively, Gdansk and Warsaw.

DOI: 10.1021/jacs.9b05895
J. Am. Chem. Soc. 2019, 141, 17142−17149

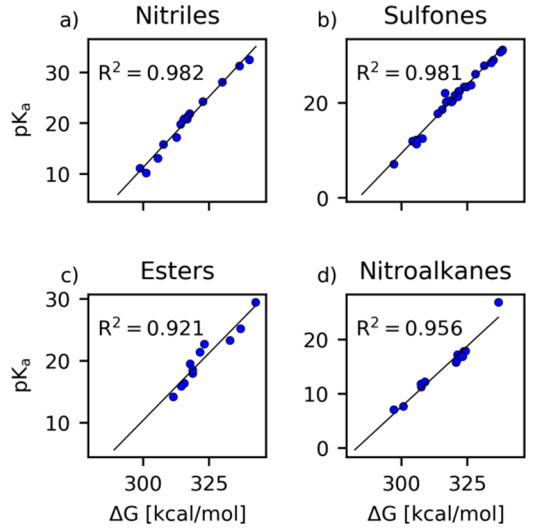


Figure 3. Correlation between calculated proton dissociation free energies and experimental pK_a's for four series of compounds: (a) nitriles (R-CN), (b) sulfones (R-SO_2Ph), (c) alkyl esters (R-COOR), and (d) nitro compounds (R-NO_2).

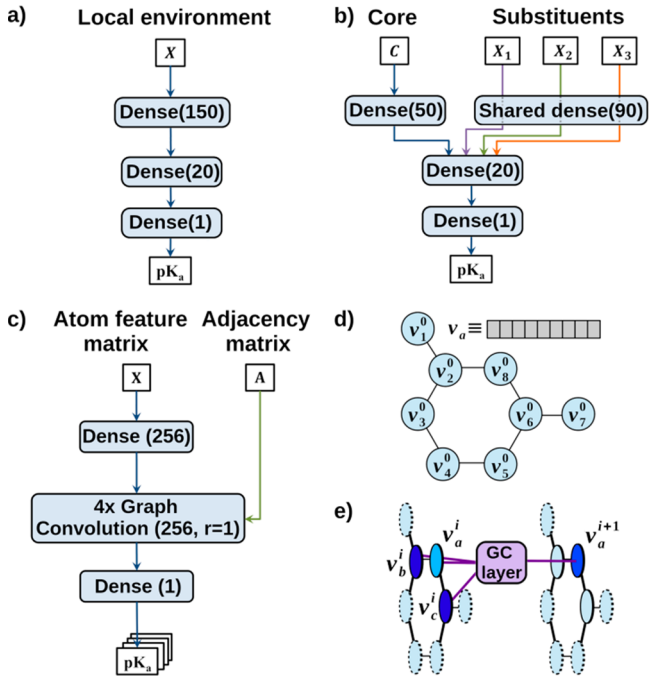


Figure 4. Schemes of neural network, NN, architectures used in the study and illustration of graph convolution. Rounded blocks denote NN layers, and the numbers in parentheses indicate the number of neurons in each layer. Rectangles denote NN inputs and output. (a) Stack of dense layers: each neuron has its inputs connected to all the neurons in the previous layer. Here, a single fingerprint vector describing the local environment of C−H acid serves as an input. (b) Multiple-input network with a shared layer. This NN takes four inputs: feature vector C describing acidic carbon and three more feature vectors X_1, X_2, X_3 characterizing each of its three possible substituents. The shared layer applies exactly the same operation (defined by the set of weights) to each of X_i vectors. The results are then combined to produce pK_a value of the considered atom. (c) Graph convolutional neural network (GCNN). The molecular graph is provided to the network as the atom feature matrix X and the adjacency matrix A. The first dense layer is intended for initial preprocessing of each atom vector (row of the X matrix), with the set of weights shared among atoms. Atomic features are then processed by graph convolution layers, analyzing the atomic neighborhoods (defined by A) up to radius r (cf. details in panels d and e). Note that this NN, in contrast to the previously described networks, produces a vector of "atomic" pK_a's rather than a single value, since the whole molecule is processed simultaneously. (d) A schematic example of a molecular graph. Each node (representing an atom) is assigned a vector of atomic descriptors (see text). Here, we denote such a vector of atom at state i as v_a^i. Connections between nodes/atoms are described by a connectivity matrix (not shown). (e) Graph convolution (GC) layer computes new feature vector v_a^{i+1} of atom a based on its previous state v_a^i and the previous states of its neighbors (here, for illustration, just nearest neighbors, v_b^i and v_c^i; the corresponding neighborhood radius r = 1). Each atom is processed with exactly the same set of weights in the layer. Note that the new feature vector may have different length than the previous one (one may produce different linear combinations of vectors' elements and there may be more such combinations than input features).

Furthermore, because our extended set did not contain any aromatic C−H's, we included in it pK_a values from theoretical study of Shen and co-workers.[15] Their set consists of 194 pK_a values computed for C−H acids in various positions of 63 aromatic molecules (e.g., in nitrobenzene, one can consider three pK_a values associated with ortho, meta, and para positions). These results were obtained using LFER approach based on B3LYP/6-311++G(2df,2p) DFT calculations in implicit solvent (DMSO), taking as a reference 10 pK_a values of heterocyclic compounds, rescaled from experimental measurements in THF. The authors obtained high correlation with experimental data (R^2 = 0.94, MAE = 0.77), which suggests that the quality of their results is comparable with that of our own QM calculations.

All in all, our training set—the composition of which is summarized in the SI, Table S6—comprised 822 molecules with accurate pK_a values. This set was also structurally diverse, as evidenced by the average Tanimoto similarity (based on comparisons of all possible pairs of molecules' ECFP4 fingerprints[17a]) being only 0.095 for the entire set and not more than ∼0.2−0.4 even within the homologous series of compounds in Figure 3 (see also SI, Table S7).

2.2. Training of Neural Networks. The combined set discussed above was used for the training of artificial neural networks (NNs) based on three types of architectures: (i) stack of dense layers, (ii) architecture with shared layers, and (iii) architecture based on graph convolution (GCNN).[11a,b] For each one, we examined several vectorization schemes (that is, ways of numerical representation of molecular and atomic information) as well as NN hyperparameters (like number of neurons per layer, number of layers, optimizer parameters, and so on). All error values reported thereafter come from 10% held-out test set. The NNs (i) and (ii) take as input description of one proton-donating atom and return only one pK_a value, whereas GCNN takes description of the whole molecule as input and returns pK_a values for all its atoms.

(i) Since stack of dense layers (the most basic and common architecture; Figure 4a) requires input vector of fixed size,[16] we chose as this input the fingerprint representation of the environment of the proton-donating fragment (within topological distance of 2; increasing this value does not improve the network's performance; see SI, Section S4). With standard ECFP4 fingerprints,[17a] the MAE was 4.5 pK_a units. However, when mol2vec[17b] was used to embed the ECFP4

DOI: 10.1021/jacs.9b05895
J. Am. Chem. Soc. 2019, 141, 17142−17149

fingerprint representation into a new vector constructed via a word embedding technique[17c] known from natural language processing, the MAE improved to 3.0 pK_a units. This value dropped even further, to 2.0, when the model was supplemented with an additional input describing the whole molecule (that is, an ECFP4 fingerprint of the whole molecule embedded with mol2vec). However, such a model—provided with information about functional groups that might be quite distant in the molecule—did not perform exceedingly well (accuracy ∼80%) when tested on the set of ∼13 000 reactions involving molecules larger than in our 822 set (cf. section 2.3 below). For such molecules, the model might be learning spurious/"confusing" information about distant and irrelevant functionalities.

(ii) Since fingerprint representation provides only topological information about atom environment (and neglects, e.g., the electronic properties of atoms or chemical groups), we also constructed a NN utilizing some chemically relevant descriptors, like Hammett constants or atom electronegativities. This model was meant to separately preprocess information about an atom of interest (given by ECFP4 fingerprint) and its substituents (characterized by Hammett constants and other descriptors) to ultimately combine the results through a common layer and compute the pK_a value (see Figure 4b). Unfortunately, this shared-layer architecture did not perform well and achieved MAE of 4.5 pK_a units. Further supplementation with global features (see previous point) reduces MAE to 3.5, which is still is worse than for the stack of dense layers.

(iii) Ultimately, we decided to implement the so-called graph convolutional neural network (GCNNs;[11] Figure 4c-e) operating on a molecular graph and algebraically represented by the atom feature matrix $N_{atoms}\times N_{features}$ and the adjacency matrix $N_{atoms}\times N_{atoms}$, where N_{atoms} is the number of atoms in the molecule and $N_{features}$ is the number of descriptors assigned to each atom (the length of the vector v_a in Figure 4d); note that only the latter number is involved in the GCNN's definition. These matrices are used to process feature vectors of individual atoms along with their neighborhood to provide new vectors of features (Figure 4e). Importantly, these new vectors effectively capture the chemical (as opposed to only topological) properties of atomic neighborhoods. Another important characteristic of this convolutional step is that all atoms within a molecule are processed in the same manner, with the same set of weights in the convolutional layer. Recently, this process has been generalized to a "message passing" framework, whereby the convolutional step corresponds to the mutual exchange of information ("messages") between connected atoms.[11c] Both GCNNs and neural-message passing neural networks (MPNNs) have been successfully applied to predict various molecular properties,[11c,d] including quantum chemical quantities such as the HOMO−LUMO gap or the dipole moment, achieving very high accuracy. However, these applications involve gathering information from all atoms in the molecule to produce global, molecule-wide quantities. On the other hand, since a similar graph convolutional framework has also been used to label individual nodes in graphs,[11b] we surmised the method would be well-suited to the prediction of atom-specific quantities such as pK_a's.

As features describing individual atoms in GCNN, we used atomic number, number of valence electrons, hybridization (different integers assigned to sp^1, sp^2, sp^3, sp^3d, and sp^3d^2), aromaticity (0 or 1), number of attached H atoms, Pauling's electronegativity, contribution to the solvent accessible area, and Gasteiger partial charge.[18] We did not include the bond-order information because, as we verified, it did not improve the results, likely because the information it provided was redundant (i.e., bond order can be deduced from atom adjacency, hybridization, and number of hydrogens attached to each heavy atom).

We tested two kinds of convolution layers: local addition layer and Chebyshev polynomial layer.[11a] The former basically constructs a linear combination from features of a given atom and its nearest neighbors, and then applies non-linear activation function to the elements of the resulting vector. The latter acts in a similar fashion but uses a Chebyshev polynomial of adjacency matrix to provide weights of vectors in the linear combination. The degree of the polynomial dictates the size of the neighborhood included in the convolution (linear, nearest neighbors; quadratic, atoms up to two bonds away; cubic, atoms up to three bonds away; etc.). Such a construction, having its justification in the spectral graph theory, was proposed by Defferrard and co-workers, yet for a different type of graph-structured data.[11a] In our tests, however, the best model involving Chebyshev layers achieved MAE of 2.8 (see SI, Table S8). We obtained better results with a GCNN model using four local layers (Figure 4c). For this model, MAE was 2.2 pK_a (2.1 pK_a in cross-validation, Figure 2), and the average computation time per molecule was ∼20 ms.

We make two remarks regarding these results. First, because GCNN analyzes neighborhood of each atom only to a certain, predefined degree (dependent on the construction of convolutional layer and the number of such layers in the network; the effective size of neighborhood processed by the network equals the sum of neighborhood sizes processed by consecutively connected convolutional layers), the network does not "perceive" the entire molecule. The use of Gasteiger partial charges as atomic descriptors is motivated by the desire to at least partly capture longer-distance interactions. Indeed, this descriptor is crucial to the model, since its negligence increases the MAE by about 0.4 pK_a unit (elimination of other descriptors results in smaller MAE increases; see SI, Figure S1).

Second, in order to verify whether the chemical meaning of the descriptors we used is actually important, we performed adversarial controls with the same GCNN architecture but with random-number features assigned to each atom. Such an input is equivalent to providing information only about atomic connectivity (since the adjacency matrix is not changed) and composition (since atoms of the same element are assigned with the same descriptors). This model achieved MAE of 3.0 pK_a units—that is, 0.8 pK_a units worse than the "chemically aware" GCNN and close to the NN based on "local" topological information alone (cf. point (i) above). Together with tests on ∼13 000 chemical reactions (cf. section 2.3), these results indicate that (1) graph-based NNs can outperform the fingerprint-based approach and also (2) that they function optimally when based on chemically meaningful

DOI: 10.1021/jacs.9b05895
J. Am. Chem. Soc. 2019, 141, 17142−17149

descriptors, whose importance we have seen when considering other chemical AI problems.[12]

2.3. Large-Scale Tests and Comparisons. To evaluate the GCNN model's performance in realistic chemical examples, we collected 12 873 pK_a-controlled reactions from the "high impact chemical journal" collection (https://github.com/jshmjs45/data_for_chem). This dataset is diverse in terms of both reaction types (aldol-type additions, Michael reactions, Claisen-type condensations, alkylations at position α to EWG groups) as well as the reaction products (average Tanimoto coefficient similarity between all reaction products equal to 0.147). On this dataset, our model correctly predicted the reacting site of C−H acid with 90.5% accuracy. As can be seen in Figure 5, the error rate decreased rapidly with the absolute difference of pK_a between the two most acidic sites.

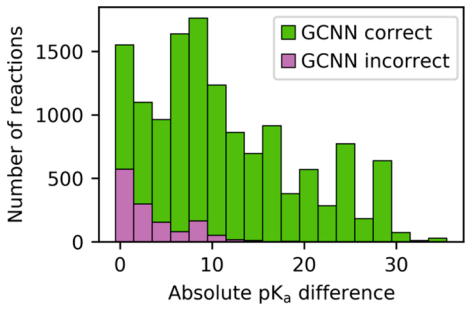


Figure 5. Histograms of correct (green bars) vs incorrect (violet bars) predictions on the set of 12 873 pK_a-controlled reactions for our GCNN model.

It is also instructive to compare our results with the predictions offered by the Weisfeiler−Lehman (W-L) GCNN network recently developed by Coley et al. to predict reactions' outcomes.[11e] This network was trained on the set of ca. 400 000 reactions from the United States Patent and Trademark Office (USPTO) set, among which some 6500 involve C−H acids. Since this network's predictions might depend on the choice of reagents (here, base), we extracted the names of the base from the textual description in the 12 873 database records wherever it was possible; for the records missing this information (~70% of cases), we used a "default" base instead, chosen to be either synthetically popular KH or LDA. The solvent was taken as THF as it is a common solvent used for C−H acid deprotonation. We then inspected whether the network predicted among any of its suggestions the experimentally observed product. However, regardless of the choice of the "default" base, the W-L network offered accuracy of only ~50% (50.9% for base = KH and 51.6% for LDA). Even with the relaxed criterion of taking the better of KH or LDA results, the accuracy was still only 53.3%. These results resonate with our previous findings that NNs yield significantly more accurate results when trained on high-quality datasets relevant to specific chemical sub-problems (cycloadditions in ref 12, pK_a here) rather than "globally" on reaction datasets that are large in size but dominated/biased by simple reaction types.[19]

2.4. Accounting for Other Factors Influencing Reaction Outcomes. In addition to thermodynamic factors discussed so far, the choice of the C−H site to be deprotonated can be affected by several other factors: (1) formation of a dianion prior to the addition of an electrophile; (2) deprotonation controlled by pre-coordination of base; (3) diminished reactivity due to formation of strained intermediates; and (4) enolization of ketones under equilibrating conditions giving thermodynamic enolates. In synthetic practice, these factors can be introduced intentionally (and effectively) to functionalize a less acidic reaction site.

When more than one equivalent of base is used, a dianion can be generated[20a] in which the less acidic site is more nucleophilic (cf. extensive studies by Mayr[20b−f]) and reacts preferentially with an electrophile. To locate this site, our web app (https://pka.allchemy.net/) first calculates the most acidic site, deprotonates it, and then recalculates the pK_a's of the remaining positions (see examples in Figure 6a−c and in SI, Figure S3).

Within group (2), a Lewis-acidic metal (usually, lithium) cation coordinates to the Lewis-basic directing group and the deprotonation step occurs in an intramolecular fashion and always in close proximity to this group (see[21a] also scheme in SI Figure S4a). To make our model applicable to such cases and also to situations when "competing" directing groups are present, we augmented it with a set of heuristics quantifying the groups' directing powers[21b,c] as either strong or moderate (see refs 21a−d and SI Figure S4b,c). In the example in Figure 6d,e, our algorithm uses these heuristics to predict correctly that chlorocarbamate used in Snieckus' synthesis of Ochratoxin A[21e] is selectively deprotonated *ortho*- to the more powerful carbamate group (dark green) rather than close to the weaker Cl atom (light green). We note that although the coordination mechanism is most commonly used during deprotonation of aromatic substrates, it can also operate during deprotonation of vinylic or aliphatic H's, as illustrated in Figure 6f,g whereby deprotonation of hydrazone during the Shapiro reaction[21f] is controlled by the configuration of C=N bond rather than acidity of adjacent protons and occurs *syn* to *N*-Ts moiety.

Within group (3), we considered deprotonation of bicyclic substrates or small-ring carbonyl compounds. Here, deprotonation is very often disfavored because abstraction of bridgehead hydrogen leads to the formation of strained, so-called non-conventional anti-Bredt anions.[22a−c] Such effects were studied in compounds having [2.2.1] framework (e.g., camphenilone in[22d,e]) and, more recently, in less strained nemorosone/clusianone/hyperforin-type natural products containing carbobicyclo[3.3.1] framework. In several cases,[22a,f−h] direct functionalization of this scaffold failed or was limited in terms of applicable electrophiles because the generated bridgehead anion showed unexpectedly low reactivity. Such situations required indirect solutions relying on the metalation of bridgehead tertiary iodide. In our software, the user is warned if bridgehead CH positions in [3.3.1] and more strained [1.1.1], [2.1.1], [2.2.1], [2.2.2], [3.1.1], and [3.2.1] systems are suggested, based solely on the pK_a, as potential deprotonation sites (Figure 6h,i).

Finally, group (4) encompasses functionalization of ketones under thermodynamic conditions. Here, the equilibrating conditions applied during generation of enolates[23a] result in functionalization of the less acidic, more substituted and nucleophilic[23b] reaction site. Classic examples include Danishefsky's synthesis of Jiadifenin[23c] or Baran's synthesis of Maoecrystal V[23d] (see SI, Figure S5). Unfortunately, we were not able to identify general-scope heuristics that would guide selection of specific reaction conditions leading to selective formation of such thermodynamic enolates. The solution to this problem might require higher-end QM

DOI: 10.1021/jacs.9b05895
J. Am. Chem. Soc. 2019, 141, 17142−17149

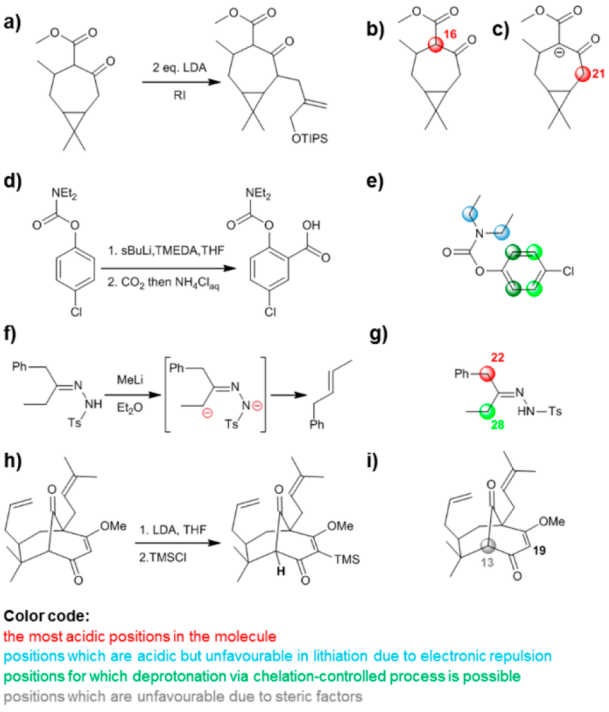


Figure 6. Extensions of the pK_a prediction model. (a–c) Accounting for the formation of a dianion upon addition of two base equivalents. Panel a illustrates alkylation of a cyclic ketoester from Funk's synthesis of the Ingenane ring system.[20g] Panels b and c show the results of Allchemy's two-step GCNN calculation whereby the most acidic site is first identified (red circle, pK_a = 16). When this position is deprotonated and a monoanion is created, the algorithm recalculates pK_a's of remaining positions and identifies site matching the experimental outcome (red circle, pK_a = 21). (d, e) Functionalization of chlorocarbamate in the total synthesis of Ochratoxin is controlled by the Lewis-basic carbamate group, which is a stronger directing group than the chlorine atom. Panel d gives the experimental result from ref 21e, and panel e shows the result of our web app in which the equivalent reaction sites *ortho* to the more powerful carbamate group are marked by dark green circles. (f, g) Coordination-controlled deprotonation in Shapiro reaction. Although benzylic protons are more acidic (as evidenced by deuterium studies), the reaction reported in ref 21f occurs exclusively *syn* to the *N*-Ts moiety due to *N*-coordination. The algorithm marks this position with a green circle while also indicating that it has higher inherent pK_a than the position marked by the red circle. (h, i) Functionalization of strained [3.3.1] framework. Direct functionalization of 1,3-diketone in Danishefsky's synthesis of nemorosone[22a] was not possible. Instead, deprotonation and silylation of the less acidic vinyl ether was observed. The algorithm marks the bridgehead position with a gray halo indicating that its deprotonation is problematic, despite apparently low calculated pK_a. The color coding used in panels b, c, e, g, and i matches that used in the https://pka.allchemy.net web app. See also SI Figures S3 and S4d for additional examples and actual screenshots.

calculations or ML on data with detailed experimental procedures, and is thus left for future research.

3. CONCLUSIONS

In summary, we used GCNNs to develop a pK_a predictor whose accuracy approaches that of QM-based models while the speed is several orders of magnitude faster. This work evidences that neural networks based on graph convolution and chemically aware descriptors taking into account atomic environments can be useful in predicting not only molecule-wide properties (as had been successfully demonstrated before[11c,d]) but also atom-specific characteristics. Above all, we hope that the web app we make available at https://pka.allchemy.net will become a useful resource assisting chemists in their day-to-day synthesis design tasks.

■ ASSOCIATED CONTENT

Supporting Information

The Supporting Information is available free of charge on the ACS Publications website at DOI: 10.1021/jacs.9b05895.

Additional theoretical details: overview of other methods for pK_a prediction, quantum-chemical benchmark calculations, details on the composition of the training set and of neural networks' setup, and description of transfer learning experiments and of other factors influencing reaction outcomes (PDF)

■ AUTHOR INFORMATION

Corresponding Author

*nanogrzybowski@gmail.com

ORCID

Bartosz A. Grzybowski: 0000-0001-6613-4261

Author Contributions

⊥R.R. and W.B. contributed equally.

Notes

The authors declare the following competing financial interest(s): While the webapp of the pKa predictor is made freely available to the community, B.A.G., R.R. and W.B. declare financial interest in Allchemy, Inc.

■ ACKNOWLEDGMENTS

This work was supported by Allchemy, Inc. R.R. acknowledges partial personal support from the National Science Center, NCN, Poland (award Fuga #2016/20/S/ST5/00361), and B.A.G., partial personal support from the Institute for Basic Science, Korea (award IBS-R020-D1). DFT calculations were carried out at the Academic Computer Centre in Gdańsk (TASK) and at the Interdisciplinary Centre for Mathematical and Computational Modelling (ICM) University of Warsaw.

■ REFERENCES

(1) Krieger, E.; Dunbrack, R. L.; Hooft, R. W. W.; Krieger, B. *Assignment of Protonation States in Proteins and Ligands: Combining pK_a Prediction with Hydrogen Bonding Network Optimization*; Springer: New York, 2012; pp 405–421.

(2) (a) Fraczkiewicz, R.; Lobell, M.; Göller, A. H.; Krenz, U.; Schoenneis, R.; Clark, R. D.; Hillisch, A. Best of Both Worlds: Combining Pharma Data and State of the Art Modeling Technology To Improve in Silico pK_a Prediction. *J. Chem. Inf. Model.* **2015**, *55*, 389–397. (b) Manallack, D. T. The pK_a Distribution of Drugs: Application to Drug Discovery. *Perspect. Med. Chem.* **2007**, *1*, 25–38. (c) Wan, H.; Ulander, J. High-Throughput pK_a Screening and Prediction Amenable for ADME Profiling. *Expert Opin. Drug Metab. Toxicol.* **2006**, *2*, 139–155. (d) Charifson, P. S.; Walters, W. P. Acidic and Basic Drugs in Medicinal Chemistry: A Perspective. *J. Med. Chem.* **2014**, *57*, 9701–9717.

(3) (a) Druais, V.; Hall, M. J.; Corsi, C.; Wendeborn, S. V.; Meyer, C.; Cossy, J. A Convergent Approach Toward the C1–C11 Subunit of Phoslactomycins and Formal Synthesis of Phoslactomycin B. *Org. Lett.* **2009**, *11* (4), 935–938. (b) Mattson, R. J.; Sloan, C. P.; Lockhart, C. C.; Catt, J. D.; Gao, Q.; Huang, S. Ortho-Directed Lithiation of 3,4-(Alkylenedioxy)Halobenzenes with LDA and LiTMP. The First Ortho Lithiation of an Iodobenzene. *J. Org.*

DOI: 10.1021/jacs.9b05895
J. Am. Chem. Soc. 2019, 141, 17142–17149

Chem. **1999**, *64* (21), 8004−8007. (c) Horn, E. J.; Silverston, J. S.; Vanderwal, C. D. A Failed Late-Stage Epimerization Thwarts an Approach to Ineleganolide. *J. Org. Chem.* **2016**, *81* (5), 1819−1838.
(4) Bordwell pK_a Table (Acidity in DMSO), http://www.chem.wisc.edu/areas/reich/pkatable/ (accessed Apr 8, 2019).
(5) Liao, C.; Nicklaus, M. C. Comparison of Nine Programs Predicting pK_a Values of Pharmaceutical Substances. *J. Chem. Inf. Model.* **2009**, *49*, 2801−2812.
(6) Bochevarov, A. D.; Watson, M. A.; Greenwood, J. R.; Philipp, D. M. Multiconformation, Density Functional Theory-Based pK_a Prediction in Application to Large, Flexible Organic Molecules with Diverse Functional Groups. *J. Chem. Theory Comput.* **2016**, *12* (12), 6001−6019.
(7) (a) Liptak, M. D.; Shields, G. C. Accurate pK_a Calculations for Carboxylic Acids Using Complete Basis Set and Gaussian-n Models Combined with CPCM Continuum Solvation Methods. *J. Am. Chem. Soc.* **2001**, *123*, 7314−7319. (b) Yu, H. S.; Watson, M. A.; Bochevarov, A. D. Weighted Averaging Scheme and Local Atomic Descriptor for pK_a Prediction Based on Density Functional Theory. *J. Chem. Inf. Model.* **2018**, *58*, 271−286. (c) Car, R.; Parrinello, M. Unified Approach for Molecular Dynamics and Density Functional Theory. *Phys. Rev. Lett.* **1985**, *55*, 2471−2474. (d) Marx, D.; Hutter, J. *Ab Initio Molecular Dynamics: Basic Theory and Advanced Methods*; Cambridge University Press: Cambridge, 2009. (e) Tummanapelli, A. K.; Vasudevan, S. Ab Initio Molecular Dynamics Simulations of Amino Acids in Aqueous Solutions: Estimating pK_a Values from Metadynamics Sampling. *J. Phys. Chem. B* **2015**, *119*, 12249−12255 and citations therein . (f) Zhang, S.; Baker, J.; Pulay, P. A Reliable and Efficient First Principles-Based Method for Predicting pK_a Values. 1. Methodology. *J. Phys. Chem. A* **2010**, *114*, 425−431. (g) Klicić, J. J.; Friesner, R. A.; Liu, S.-Y.; Guida, W. C. Accurate Prediction of Acidity Constants in Aqueous Solution via Density Functional Theory and Self-Consistent Reaction Field Methods. *J. Phys. Chem. A* **2002**, *106*, 1327−1335. (h) Zuo, C.-S.; Wiest, O.; Wu, Y.-D. Parameterization and Validation of Solvation Corrected Atomic Radii. *J. Phys. Chem. A* **2009**, *113*, 12028−12034.
(8) (a) Perrin, D. D.; Dempsey, B.; Serjeant, E. P. *pK_a prediction for organic acids and bases*; Chapman and Hall: London, 1981. (b) Shields, G. C.; Seybold, P. G. *Computational Approaches for the Prediction of pK_a Values*; CRC Press: Boca Raton, 2014. (c) Svobodová Vařeková, R.; Geidl, S.; Ionescu, C.-M.; Skřehota, O.; Bouchal, T.; Sehnal, D.; Abagyan, R.; Koča, J. Predicting pK_a Values from EEM Atomic Charges. *J. Cheminf.* **2013**, *5*, 18. (d) Yu, H.; Kühne, R.; Ebert, R.-U.; Schüürmann, G. Comparative Analysis of QSAR Models for Predicting pK_a of Organic Oxygen Acids and Nitrogen Bases from Molecular Structure. *J. Chem. Inf. Model.* **2010**, *50*, 1949−1960. (e) Ding, F.; Smith, J. M.; Wang, H. First-Principles Calculation of pK_a Values for Organic Acids in Nonaqueous Solution. *J. Org. Chem.* **2009**, *74*, 2679−2691.
(9) (a) Klucznik, T.; Mikulak-Klucznik, B.; McCormack, M. P.; Lima, H.; Szymkuć, S.; Bhowmick, M.; Molga, K.; Zhou, Y.; Rickershauser, L.; Gajewska, E. P.; et al. Efficient Syntheses of Diverse, Medicinally Relevant Targets Planned by Computer and Executed in the Laboratory. *Chem.* **2018**, *4*, 522−532. (b) Coley, C. W.; Rogers, L.; Green, W. H.; Jensen, K. F. Computer-Assisted Retrosynthesis Based on Molecular Similarity. *ACS Cent. Sci.* **2017**, *3*, 1237−1245. (c) Segler, M. H. S.; Preuss, M.; Waller, M. P. Planning Chemical Syntheses with Deep Neural Networks and Symbolic AI. *Nature* **2018**, *555*, 604−610.
(10) (a) Greenwood, J. R.; Calkins, D.; Sullivan, A. P.; Shelley, J. C. Towards the Comprehensive, Rapid, and Accurate Prediction of the Favorable Tautomeric States of Drug-like Molecules in Aqueous Solution. *J. Comput.-Aided Mol. Des.* **2010**, *24*, 591−604. (b) *Epik: Rapid and robust pKa predictions*; Schrödinger, LLC, New York, 2019; https://www.schrodinger.com/epik.
(11) (a) Defferrard, M.; Bresson, X.; Vandergheynst, P. Advances in Neural Information Processing Systems: 30th Annual Conference on Neural Information Processing Systems, 2016. (b) Kipf, T. N.; Welling, M. Semi-Supervised Classification with Graph Convolutional Networks. *arXiv Preprint* arXiv:1609.02907 [cs.LG], 2016. (c) Gilmer, J.; Schoenholz, S. S.; Riley, P. F.; Vinyals, O.; Dahl, G. E. Neural Message Passing for Quantum Chemistry. *Proc. Machine Learning Res.* **2017**, 1263−1272. (d) Faber, F. A.; Hutchison, L.; Huang, B.; Gilmer, J.; Schoenholz, S. S.; Dahl, G. E.; Vinyals, O.; Kearnes, S.; Riley, P. F.; von Lilienfeld, O. A. Prediction Errors of Molecular Machine Learning Models Lower than Hybrid DFT Error. *J. Chem. Theory Comput.* **2017**, *13*, 5255−5264. (e) Coley, C. W.; Jin, W.; Rogers, L.; Jamison, T. F.; Jaakkola, T. S.; Green, W. H.; Barzilay, R.; Jensen, K. F. A graph-convolutional neural network model for the prediction of chemical reactivity. *Chem. Sci.* **2019**, *10*, 370−377. (f) Yang, K.; Swanson, K.; Jin, W.; Coley, C.; Eiden, P.; Gao, H.; Guzman-Perez, A.; Hopper, T.; Kelley, B. P.; Mathea, M.; Palmer, A.; Settels, V.; Jaakkola, T. S.; Jensen, K. F.; Barzilay, R. Are Learned Molecular Representations Ready for Prime Time?. *ChemRxiv Preprint*, 2019, DOI: 10.26434/chemrxiv.7940594.v1. (g) Duvenaud, D.; Maclaurin, D.; Aguilera-Iparraguirre, J.; Gomez-Bombarelli, R.; Hirzel, T.; Aspuru-Guzik, A.; Adams, R. P. Convolutional Networks on Graphs for Learning Molecular Fingerprints. Proceedings of Advances in Neural Information Processing Systems 28 (NIPS 2015), Montreal, Canada, Dec 7−12, 2015; pp 2215−2223 (h) Kearnes, S.; McCloskey, K.; Berndl, M.; Pande, V.; Riley, P. Molecular graph convolutions: moving beyond fingerprints. *J. Comput.-Aided Mol. Des.* **2016**, *30*, 595−608.
(12) Beker, W.; Gajewska, E. P.; Badowski, T.; Grzybowski, B. A. Prediction of Major Regio-, Site-, and Diastereoisomers in Diels-Alder Reactions by Using Machine-Learning: The Importance of Physically Meaningful Descriptors. *Angew. Chem., Int. Ed.* **2019**, *58*, 4515−4519.
(13) Streitwieser, A.; Wang, D. Z.; Stratakis, M.; Facchetti, A.; Gareyev, R.; Abbotto, A.; Krom, J. A.; Kilway, K. V. Extended Lithium Ion Pair Indicator Scale in Tetrahydrofuran. *Can. J. Chem.* **1998**, *76*, 765−769.
(14) (a) Halgren, T. A. Merck Molecular Force Field. I. Basis, Form, Scope, Parameterization, and Performance of MMFF94. *J. Comput. Chem.* **1996**, *17*, 490−519. (b) Rappe, A. K.; Casewit, C. J.; Colwell, K. S.; Goddard, W. A.; Skiff, W. M. UFF, a Full Periodic Table Force Field for Molecular Mechanics and Molecular Dynamics Simulations. *J. Am. Chem. Soc.* **1992**, *114*, 10024−10035.
(15) Shen, K.; Fu, Y.; Li, J.-N.; Liu, L.; Guo, Q.-X. What Are the pK_a Values of C−H Bonds in Aromatic Heterocyclic Compounds in DMSO? *Tetrahedron* **2007**, *63*, 1568−1576.
(16) Bianchini, M.; Maggini, M.; Jain, L. C. *Handbook on Neural Information Processing*; Springer, 2013.
(17) (a) Rogers, D.; Hahn, M. Extended-Connectivity Fingerprints. *J. Chem. Inf. Model.* **2010**, *50*, 742−754. (b) Jaeger, S.; Fulle, S.; Turk, S. Mol2vec: Unsupervised Machine Learning Approach with Chemical Intuition. *J. Chem. Inf. Model.* **2018**, *58*, 27−35. (c) Mikolov, T.; Chen, K.; Corrado, G.; Dean, J. Efficient Estimation of Word Representations in Vector Space. *arXiv Preprint* arXiv:1301.3781 [cs.CL], 2013.
(18) Gasteiger, J.; Marsili, M. Iterative Partial Equalization of Orbital Electronegativity—a Rapid Access to Atomic Charges. *Tetrahedron* **1980**, *36*, 3219−3228.
(19) For discussion of reaction types within the USPTO set, see: Jaworski, W.; Szymkuć, S.; Mikulak-Klucznik, B.; Piecuch, K.; Klucznik, T.; Kaźmierowski, M.; Rydzewski, J.; Gambin, A.; Grzybowski, B. A. *Nat. Commun.* **2019**, *10*, 1434.
(20) (a) Huckin, S. N.; Weiler, L. Alkylation of Dianions of Beta-Keto Esters. *J. Am. Chem. Soc.* **1974**, *96*, 1082−1087. (b) Mayr, H.; Patz, M. Scales of Nucleophilicity and Electrophilicity: A System for Ordering Polar Organic and Organometallic Reactions. *Angew. Chem., Int. Ed. Engl.* **1994**, *33*, 938−957. (c) Mayr, H.; Ofial, A. R. Do General Nucleophilicity Scales Exist? *J. Phys. Org. Chem.* **2008**, *21*, 584−595. (d) Mayr, H.; Ofial, A. R. Kinetics of Electrophile-Nucleophile Combinations: A General Approach to Polar Organic Reactivity. *Pure Appl. Chem.* **2005**, *77*, 1807−1821. (e) Kaumanns, O.; Appel, R.; Lemek, T.; Seeliger, F.; Mayr, H. Nucleophilicities of the Anions of Arylacetonitriles and Arylpropionitriles in Dimethyl Sulfoxide. *J. Org. Chem.* **2009**, *74*, 75−81. (f) Mayr's Database Of

DOI: 10.1021/jacs.9b05895
J. Am. Chem. Soc. 2019, 141, 17142−17149

Reactivity Parameters, https://www.cup.lmu.de/oc/mayr/reaktionsdatenbank/. (g) Funk, R. L.; Olmstead, T. A.; Parvez, M.; Stallman, J. B. Stereoselective Construction of the Complete Ingenane Ring System. *J. Org. Chem.* **1993**, *58*, 5873−5875.

(21) (a) Snieckus, V. Directed Ortho Metalation. Tertiary Amide and *O*-Carbamate Directors in Synthetic Strategies for Polysubstituted Aromatics. *Chem. Rev.* **1990**, *90*, 879−933. (b) Miah, M. A. J.; Sibi, M. P.; Chattopadhyay, S.; Familoni, O. B.; Snieckus, V. Directed Ortho-Metalation of Aryl Amides, *O*-Carbamates, and Methoxymethoxy Systems: Directed Metalation Group Competition and Cooperation. *Eur. J. Org. Chem.* **2018**, *2018*, 447−454. (c) Clayden, J. Regioselective Synthesis of Organolithiums by Deprotonation. *Organolithiums: Selectivity for Synthesis*; Elsevier, 2002; pp 9−109. (d) Miah, M. A. J.; Sibi, M. P.; Chattopadhyay, S.; Familoni, O. B.; Snieckus, V. Directed Ortho-Metalation of Aryl Amides, O-Carbamates, and Methoxymethoxy Systems: Directed Metalation Group Competition and Cooperation. *Eur. J. Org. Chem.* **2018**, *2018*, 447−454. (e) Sibi, M. P.; Chattopadhyay, S.; Dankwardt, J. W.; Snieckus, V. Combinational O-Aryl Carbamate and Benzamide Directed Ortho Metalation Reactions. Synthesis of Ochratoxin A and Ochratoxin B. *J. Am. Chem. Soc.* **1985**, *107*, 6312−6315. (f) Shapiro, R. H.; Lipton, M. F.; Kolonko, K. J.; Buswell, R. L.; Capuano, L. A. Tosylhydrazones and Alkyllithium Reagents: More on the Regiospecificity of the Reaction and the Trapping of Three Intermediates. *Tetrahedron Lett.* **1975**, *16*, 1811−1814.

(22) (a) Tsukano, C.; Siegel, D. R.; Danishefsky, S. J. Differentiation of Nonconventional "Carbanions" − The Total Synthesis of Nemorosone and Clusianone. *Angew. Chem., Int. Ed.* **2007**, *46*, 8840−8844. (b) Shiner, C. S.; Berks, A. H.; Fisher, A. M. Metalation of Nonenolizable Ketones and Aldehydes by Lithium Dialkylamide Bases. *J. Am. Chem. Soc.* **1988**, *110*, 957−958. (c) Hayes, C. J.; Simpkins, N. S.; Kirk, D. T.; Mitchell, L.; Baudoux, J.; Blake, A. J.; Wilson, C. Bridgehead Lithiation-Substitution of Bridged Ketones, Lactones, Lactams, and Imides: Experimental Observations and Computational Insights. *J. Am. Chem. Soc.* **2009**, *131*, 8196−8210. (d) Nickon, A.; Lambert, J. L. S. J. Homoenolate Anions. *J. Am. Chem. Soc.* **1962**, *84*, 4604−4605. (e) Shiner, C. S.; Berks, A. H.; Fisher, A. M. Metalation of Nonenolizable Ketones and Aldehydes by Lithium Dialkylamide Bases. *J. Am. Chem. Soc.* **1988**, *110*, 957−958. (f) Ahmad, N. M.; Rodeschini, V.; Simpkins, N. S.; Ward, S. E.; Blake, A. J. Synthesis of Polyprenylated Acylphloroglucinols Using Bridgehead Lithiation: The Total Synthesis of Racemic Clusianone and a Formal Synthesis of Racemic Garsubellin A. *J. Org. Chem.* **2007**, *72*, 4803−4815. (g) Bellavance, G.; Barriault, L. Modular Total Syntheses of Hyperforin, Papuaforins A, B, and C via Gold(I)-Catalyzed Carbocyclization. *J. Org. Chem.* **2018**, *83*, 7215−7230. (h) Simpkins, N.; Taylor, J.; Weller, M.; Hayes, C. Synthesis of Nemorosone via a Difficult Bridgehead Substitution Reaction. *Synlett* **2010**, *4*, 639−643.

(23) (a) Braun, M. General Methods for the Preparation of Enolates. *Modern Enolate Chemistry*; Wiley-VCH Verlag GmbH & Co. KGaA: Weinheim, Germany, 2015; pp 11−82. (b) Quesnel, Y.; Bidois-Sery, L.; Poirier, J.-M.; Duhamel, L. Highly Regioselective Alkylation at the More Hindered γ -Site of Unsymmetrical Ketones by Use of Their Potassium Enolates. A Comparative Study with Lithium Enolates. *Synlett* **1998**, *4*, 413−415. (c) Carcache, D. A.; Cho, Y. S.; Hua, Z.; Tian, Y.; Li, Y.-M.; Danishefsky, S. J. Total Synthesis of (±)-Jiadifenin and Studies Directed to Understanding Its SAR: Probing Mechanistic and Stereochemical Issues in Palladium-Mediated Allylation of Enolate-Like Structures. *J. Am. Chem. Soc.* **2006**, *128*, 1016−1022. (d) Cernijenko, A.; Risgaard, R.; Baran, P. S. 11-Step Total Synthesis of (−)-Maoecrystal V. *J. Am. Chem. Soc.* **2016**, *138*, 9425−9428.

DOI: 10.1021/jacs.9b05895
J. Am. Chem. Soc. 2019, 141, 17142−17149