Inverse molecular design of alkoxides and phenoxides for aqueous direct air capture of CO₂

Zisheng Zhang, Amanda L. Kummeth, Janny Y. Yang, and Anastassia N. Alexandrova

Aqueous direct air capture (DAC) is a key technology toward a carbon negative infrastructure. Developing sorbent molecules with water and oxygen tolerance and high CO₂ binding capacity is therefore highly desired. We analyze the CO₂ absorption chemistries on amines, alkoxides, and phenoxides with density functional theory calculations, and perform inverse molecular design of the optimal sorbent. The alkoxides and phenoxides are found to be more suitable for aqueous DAC than amines thanks to their water tolerance (lower pK_a prevents protonation by water) and capture stoichiometry of 1:1 (2:1 for amines). All three molecular systems are found to generally obey the same linear scaling relationship (LSR) between pK_a, pK_b, since both CO₂ and proton are bonded to the nucleophilic (alkoxy or amine) binding site through a majorly σ bonding orbital. Several high-performance alkoxides are proposed from the computational screening. Phenoxides have comparatively poorer correlation between pK_a, and pK_b, showing promise for optimization. We apply a genetic algorithm to search the chemical space of substituted phenoxides for the optimal sorbent. Several promising off-LSR candidates are discovered. The most promising one features bulky ortho substitutions forcing the CO₂ adduct into a perpendicular configuration with respect to the aromatic ring. In this configuration, the phenoxide binds CO₂ and a proton using different molecular orbitals, thereby decoupling the pK_a and pK_b. The pK_a−pK_b trend and off-LSR behaviors are then confirmed by experiments, validating the inverse molecular design framework. This work not only extensively studies the chemistry of the aqueous DAC, but also presents a transferrable computational workflow for understanding and optimization of other functional molecules.
design, or molecular engineering, is best used for molecular systems where the relationship between the electronic structure and desired properties is well understood. Modifications can be made to the parent molecule to tune the molecular properties based on the design principles. On the other hand, inverse design is frequently used when there is insufficient insight into the structure–activity relationship; the pool of possible candidates can be screened, or molecular design can be treated as an optimization task where evolutionary algorithms can be applied to efficiently optimize the desired property in a predefined chemical subspace. Both strategies have proven successful in optimizing functional molecules for various applications, including solar cell (10), redox flow cell (11), solar heat battery (12), molecular photocatalysis (13), and electrocatalysis (14–17).

Here, we focus on aqueous DAC and search for the optimal sorbent molecule from various molecular systems with a combination of inverse design strategies and modern computational methods. The bonding nature and energetics of CO2 binding and protonation on substituted amines, alkoxides, and phenoxides are investigated using density functional theory (DFT) calculations. We discovered that all three sorbent families generally follow a linear scaling relationship (LSR) between pK\textsubscript{CO2} and pK\textsubscript{a} that originates from the same bonding nature of protons and CO2 to the binding site. However, the correlation between these two properties in phenoxides is weaker. To exploit the off-LSR behavior of phenoxides, a genetic algorithm (GA) searcher combined with semiempirical quantum mechanical (SQM) calculations is performed to efficiently search for air- and water-stable alkoxide species with optimal CO2 binding capabilities (search direction perpendicular to the LSR). The top-scoring candidates are further refined with DFT calculations and analyzed computationally to provide insights and design principles based on the search results. The data generated from the GA search can be further used to train predictive machine learning (ML) models for low-cost prediction of molecular orbital (MO) energy levels and binding free energies. For validation, several simple synthetically accessible substituted phenoxide and alkoxide molecules were selected to evaluate their CO2 absorption capacity. The overall pK\textsubscript{CO2}–pK\textsubscript{a} trend and off-LSR behavior of specific molecules are observed in the experiments, consistent with theoretical predictions. The inverse molecular design workflow not only provides trends and design principles, with minimal prior knowledge, but also is readily generalizable to explore and optimize other functional molecules for various applications.

**Results and Discussion**

**Amines for Aqueous DAC.** Amines are the current leading CO2 sorbent family for postcombustion capture (5, 18). Amines are usually divided into groups based on the number of hydrogens on the nitrogen, namely, primary, secondary, and tertiary. In aqueous solution, tertiary amines (R\textsubscript{1}R\textsubscript{2}R\textsubscript{3}N) hydrolyze to produce hydrxide that acts as the capture agent:

\[
\begin{align*}
R_1R_2R_3N + H_2O &\rightarrow R_1R_2R_3\text{NH}^+ + OH^- \\
CO_2 + OH^- &\rightarrow HCO_3^- \\
HCO_3^- + OH^- &\rightarrow H_2O + CO_3^{2-}.
\end{align*}
\]

In these cases, the tertiary amino group functions solely as a weak base instead of binding directly to CO2. As a result, tertiary amines are less interesting and promising from the perspective of molecular design, as substituents only affect the basicity but not the CO2 binding energetics.

CO2 capture using primary and secondary amines has been investigated extensively experimentally; however, a clear relation that could guide molecular design has not yet been proposed (19, 20). Fig. 1A shows the potential energy surface (PES) of CO2 binding on MeNH\textsubscript{2} and its deprotonated form MeNH\textsubscript{2}\textsuperscript{−} obtained from relaxed scan at B3LYP-D3(BJ)/def2-TZVP. Despite some older reports that the primary and secondary amines could capture CO2 via a zwiterionic (i.e., carboxylic acid intermediate) pathway (21, 22), our DFT calculation (Fig. 1A) shows that the neutral amino group in MeNH\textsubscript{2} could only noncovalently physisorb a CO2 molecule, as is evidenced by the N–C distance of 2.82 Å (Fig. 1B). Only after deprotonation can a primary or secondary amino group bind CO2 as a carbamate intermediate with a binding energy of −53.1 kcal/mol and an N–C bond length of 1.43 Å. Since the neutral amino groups are always more basic than water, the deprotonated amino group can only form from an autoionization process:

\[
2R_1R_2NH\rightarrow R_1R_2NH_2^+ + R_1R_2N^-.
\]

The deprotonated amine group R\textsubscript{1}R\textsubscript{2}N\textsuperscript{−} then readily captures a dissolved CO2 molecule to form carbamate:

\[
R_1R_2N^- + CO_2 \rightarrow R_1R_2NCOO^-,
\]

that in turn promotes the autoionization by consumption of the deprotonated species, until the system reaches equilibrium. Meanwhile, the R\textsubscript{1}R\textsubscript{2}NH\textsubscript{2}\textsuperscript{+} produced in the autoionization step serves no absorbing function and remains in the protonated form thereafter. Therefore, the primary and secondary amines bind CO2 in an overall stoichiometry of 2:1, i.e., one mole of amine is wasted for each mole of CO2 captured.

Since only a dilute stream of CO2 is available in CCC to achieve the optimal absorption capacity, a sorbent needs favorable thermodynamics in both autoionization and CO2 binding. To investigate the energetic trend, we perform DFT calculations on a library of monosubstituted primary and secondary amines, and Fig. 1C shows the scatter plot of their CO2 binding free energy G\textsubscript{bind} versus autoionization free energy G\textsubscript{auto–ionization}. Unfortunately, G\textsubscript{auto–ionization} lies between 39.5 and 60.9 kcal/mol, suggesting difficult autoionization and hence a low concentration of the active species for CO2 binding. In addition, no clear relation can be observed between the two reaction energies. As G\textsubscript{auto–ionization} increases, the G\textsubscript{bind} decreases until G\textsubscript{auto–ionization} reaches c.a. 52 kcal/mol and increases thereafter, forming a convex hull. Since the unsubstituted methylamine (G\textsubscript{bind} = 46.6 kcal/mol, G\textsubscript{auto–ionization} = −51.1 kcal/mol) is already located near the top of the inverted volcano, the intrinsic limit leaves little room for further optimization of both autoionization and CO2 binding.

To gain deeper insight into the energetic trend, we calculated the pK\textsubscript{A1} (R\textsubscript{1}R\textsubscript{2}R\textsubscript{3}NH\textsubscript{2} → R\textsubscript{1}R\textsubscript{2}R\textsubscript{3}N\textsuperscript{−} + R\textsubscript{1}R\textsubscript{2}H\textsuperscript{+}) and pK\textsubscript{CO2}, of the primary and secondary amines via the linear free energy relation (LFER) (23) and plotted them in Fig. 1D. Note that pK\textsubscript{A1} is usually reported as pK\textsubscript{A} due to experimental difficulty in determining pK\textsubscript{A2}. Previous experimental studies have reported a weak positive correlation between the pK\textsubscript{A1} and CO2 absorption capacity (20), which is also observed in our calculations where pK\textsubscript{A1} correlates poorly with pK\textsubscript{CO2} with a R\textsuperscript{2} of 0.22 (SI Appendix, Fig. S2A). However, pK\textsubscript{A2} is found to correlate much better with pK\textsubscript{CO2} in a linear manner, with a R\textsuperscript{2} of 0.86 (SI Appendix, Fig. S2B). The linear correlation is attributed to the similar bonding mode through which proton and CO2 act as electrophiles and bind to the nucleophilic anionic nitrogen site. The similar bonding characteristics results in coupling of the
$G_{\text{bind}}$ and $G_{\text{deprot}}$, i.e., LSR (24). The poor correlation between $pK_{\text{CO}_2}$ and $pK_{\text{a1}}$ originates in the poor correlation between $pK_{\text{a2}}$ and $pK_{\text{a1}}$ (SI Appendix, Fig. S2C). Then itrogen or deprotonated amino groups has different types of hybridization and hence is influenced to different extents by substituent effects. In addition, the first and the second proton to be attached to the anionic nitrogen is impacted by different steric effects from nearby groups. This is also the cause of the $G_{\text{bind}} - G_{\text{auto- ionization}}$ convex hull observed in Fig. 1C.

**Alkoxides for Aqueous DAC.** Despite the favorable CO$_2$ binding energetics of amines, the stoichiometry of 2:1 for the capturing agents to CO$_2$ caps the absorption capacity. The $pK_G$ of primary and secondary amines lies above 20, which makes them unsuitable for aqueous DAC due to susceptibility to protonation by water ($pK_\nu = 15.7$) and blocking of the CO$_2$ binding site. In addition, primary and secondary amines could suffer from oxidation when in contact with air. In contrast, alkoxides are a promising family of CO$_2$ capturing agents that are more O$_2$ insensitive and can directly bind CO$_2$ to form a carbonate with 1:1 stoichiometry:

$$\text{RO}^- + \text{CO}_2 \rightarrow \text{ROCO}_2^-.$$

The CO$_2$ binding to an alkoxy group has the same bonding characteristics as on a deprotonated amino group. Fig. 2A shows the two localized molecular orbitals (LMOs) with the highest contribution to the Mayer bond order between CO$_2$ and the capturing agent. It can be seen that the major bonding orbital of the amine– and alkoxy–CO$_2$ adduct both have a $\sigma$-bonding pattern, accounting for 84.7 and 78.4% of the Mayer bond order, respectively. The LMO with the second highest bonding contribution has a $\pi$-bonding pattern for both amine and alkoxides, with a minor contribution of 19.0% and 19.6%, respectively. The analysis reveals that CO$_2$ binds in a similar bonding pattern to both an alkoxy and an amine.

To explore the $pK_{\text{CO}_2} - pK_\nu$ relation of alkoxides, the binding constants are calculated using DFT and plotted in Fig. 2B for the library of monosubstituted methoxide and ethoxides, together with the datapoints of amines. The $pK_{\text{CO}_2}$ of the majority of the alkoxides lie in the range of −10 to −20, suggesting less favorable CO$_2$ binding energetics compared to amines. However, most of the alkoxides have their $pK_\nu$ below 15.7, which prevents deactivation from water protonation. This could be attributed to the weaker nucleophilicity of the O in alkoxide (Hirshfeld charge: 0.84 $e$) than the N in amine (Hirshfeld charge: −1.12 $e$). As can be expected from the chemical bonding analysis, there also exists an LSR between $pK_{\text{CO}_2}$ and $pK_\nu$ of alkoxides that is similar to the case of amines. More intriguingly, the LSRs of alkoxides and amines lie in the same straight line, with a $R^2$ of 0.97 on the combined dataset. The shared LSR demonstrates that the weakening of CO$_2$ binding on alkoxy compared to on amine is just a
compromise in exchange for water tolerance. Alkoxides are not systematically inferior to amines, and since they follow the exact same $pK_{CO_2}$–$pK_a$ curve, we can access the upper left region that is inaccessible for amines via functional group substitution. Since the $pK_a$ of ethanol ($pK_a = 16$) is slightly higher than that of water, a straightforward approach is to move toward the upper left direction along the LSR by attaching a weak electron withdrawing group (EWG) to the alkoxide. At the same time, the tradeoff must be kept small enough so that the CO$_2$ binding capability is not excessively weakened.

The top-10 water-tolerant alkoxide candidates ranked by $pK_{CO_2}$ are listed in Fig. 2C.

Optimizing Phenoxides with GA Search. The LSR provides a clear correlation along which we could tune the binding energetics; however, it also imposes an intrinsic limitation that prevents the optimization of both $pK_a$ and $pK_{CO_2}$. This is similar to the activity volcano in catalysis (25). Hence, we further extend the study to phenoxides whose distinct steric and electronic structure characteristics (aromaticity) could span a different chemical space from alkoxides. The bonding nature of CO$_2$ on phenoxide was found to be similar to that of amines and alkoxides, namely, major $\sigma$-characteristics and minor $\pi$-characteristics (Fig. 2A). A relatively low p-contribution could be attributed to the slightly out-of-plane CO$_2$ binding configuration ($156.7^\circ$ dihedral angle between the CO$_2$ plane and the benzene plane) and the lack of hyperconjugation effects by alkyl groups. Because the phenyl group is an inductive EWG, it results in a less nucleophilic O in phenoxide (Hirshfeld charge: $-0.58 \ \delta$) compared to the alkoxide case, which is why phenol is more acidic than alcohol. Therefore, phenoxides show weaker CO$_2$ binding than alkoxides or amines, in terms of both energetics and bond lengths (Fig. 1A and B). In Fig. 2B, the datapoints of mono- and disubstituted phenoxides lie in the upper left region in the $pK_{CO_2}$–$pK_a$ plot, indicating weak basicity (not prone to protonation by water) but small CO$_2$ adsorption capacity.

A closer inspection of the phenoxide datapoints could reveal a $pK_{CO_2}$–$pK_a$ distribution pattern that is distinct from that of the alkoxides. In Fig. 3A, the datapoints of alkoxides closely obey a linear correlation with an $R^2$ of 0.86. However, the datapoints of phenoxides, although roughly showing a consistent trend with the LSR of alkoxides, are much more dispersed and have a lower $R^2$ of 0.67, suggesting a weak correlation between the $pK_{CO_2}$ and $pK_a$ of phenoxides. This can be attributed to the rigidity of the phenyl group; the C–C bonds in the phenyl ring cannot rotate as freely as C–C bonds in the alkyl chains, which prevents ortho substituents from adapting to configurations where their interaction with the CO$_2$ binding site is minimized. The aromaticity of the phenyl ring also allows for para and meta substituents to influence the electronic structure of the CO$_2$ binding site via the conjugated $\pi$ system, whereas the substituent effects in alkoxides tends to dissipate beyond the $\beta$ carbon atom due to the saturated sp$^3$ hybridization. In addition, the unsubstituted phenoxide is at about the
middle of the distribution, unlike the case of alkoxides where the unsubstituted methoxide is at the lower-left end. In summary, the phenoxides have greater room for further optimization and, more importantly, the potential to break the LSR and decouple the $pK_a$ and $pK_{CO2}$.

Due to a lack of insights into the origin of such off-LSR behavior, we turn to the inverse design strategy and extend the chemical subspace from mono- and disubstitution to all possible substitutions. This would include a total of c.a. 8 million unique molecules (24 substituents, 5 sites), which is beyond the capability of brute force exhaustion. To efficiently explore the vast chemical space, we employ GA, an evolutionary algorithm that has been successfully applied to structural prediction and property optimization of molecular systems (15, 26), to search for the substituted phenoxide with minimal $\Delta G_{bind}$ on the condition that its $pK_a$ is lower than 15.74. To lower the computational cost and speed up the GA search, the SQM method GFN1-xtb, with GBSA implicit solvation, is adopted. The GFNn-xtb method has been reported to predict $pK_a$ and $pK_{CO2}$ values versus the $pK_a$ of (A) substituted alkoxides and (B) substituted phenoxides, with the $R^2$ value and the formula of the LSR fitting labeled above each plot. The green dashed line and the blue dotted line represent the data points of phenoxide and methoxide, respectively.

Ten independent GA searches are performed to avoid premature convergence in local optima, and the evolution of the lowest $\Delta G_{bind}$ is shown in SI Appendix, Fig. S4. The optimal candidate from each individual GA search outperforms the reference species by 9 to $\sim 15$ kcal/mol in terms of SQM-calculated $\Delta G_{bind}$. It can be seen from Fig. 4A that the GA not only samples sufficiently into the lower right region in the plot along the LSR but also explores the lower left region where the binding energies of the proton and $CO_2$ are decoupled. The histogram in Fig. 4A could more clearly demonstrate the sampling efficiency of the GA compared to random sampling, with 51% of the sampled candidates outperforming the reference molecule, which is a significantly higher percentage than 6% in random sampling. To verify the GA search results, DFT calculations are performed on the top-scoring candidates from the GA search at the same level of theory as in the previous sections. It is shown in Fig. 4B that all the candidates found in GA searches outperform the reference molecule, with a large portion of the datapoints distributed in the lower left off-LSR region. The datapoint with the most negative $pK_{CO2}$ is located far from the reference species and almost enters the regime of alkoxides. After filtering out the water-sensitive compounds, we rank the candidates from GA searches and show their molecular structure and binding constants in Fig. 4C.

### Data-Driven Design Principles and Predictive ML Models.

After obtaining and validating the optimal candidates from the inverse design strategy, we proceeded to study their molecular structures, aiming to understand the origin of their favorable and off-LSR energetics. A quick glance at Fig. 4C reveals the common characteristics of the top-scoring and off-LSR candidates, namely, bulky alkyl groups on the ortho positions (other molecular characteristics are discussed in SI Appendix, Note S1). This is quite unexpected since alkyl groups are neither strong EWGs nor electron donating groups, and they are usually considered to weaken binding due to steric hindrance. During inspection of the DFT-optimized geometries, we notice that the $CO_2$ binding configuration on the top-scoring candidates is quite different from on the simple substituted phenoxides. Specifically, the dihedral angle between the $CO_2$ plane and phenyl plane is c.a. 25° (noted as in-plane) for mono- and disubstituted phenoxides but 60 to $\sim 90^\circ$ (noted as perpendicular) for the top-scoring candidates with bulky ortho groups from the GA search. Moreover, as is shown in Fig. 5A, the $pK_{CO2}$ is negatively correlated with the dihedral angle, i.e., perpendicular binding configurations are more energetically favorable compared to in-plane binding configurations.

To investigate how the bulky groups at ortho positions could alter the binding energetics, we use phenoxide and 2,6-dimethylphenoxide (C2) as model systems. Note that the optimal candidate from the GA search is not used here for clarity and controlling variables. Relaxed PES scans are performed to explore the energy profile as the Dih(C-C-O-C), the dihedral angle between phenyl plane and $CO_2$ planes, rotates. The PES of the phenoxide-$CO_2$ adduct (Fig. 5B) is quite flat in the 0 to $\sim 40^\circ$ region, with the global minimum configuration at c.a. 20°. The exact in-plane 0° configuration is a local maximum due to the steric repulsion between $CO_2$ and the ortho C-H. As the Dih(C-C-O-C) increases, the phenoxide–$CO_2$ adduct gets increasingly unstable until it passes through the global maximum with respect to Dih(C-C-O-C), at $90^\circ$. However, this unstable perpendicular configuration is the global minimum configuration on the PES of the C2-$CO_2$ adduct (Fig. 5B). The steric effect of ortho methyl groups destabilizes the in-plane configuration into a local maximum. On the contrary, the perpendicular configuration is not affected as much and becomes the global minimum.

In comparison, the PES of proton binding on phenoxides is not reshaped by bulky ortho groups as occurs with $CO_2$ binding. As shown in Fig. 5C, the in-plane configuration is the global minimum for proton binding on both phenoxide and C2. The in-plane proton bindings on both molecules are contributed majorly by a $\sigma$ bonding orbital formed from H 1s and
the HOMO-1 of phenoxides (Fig. 5D). In contrast, the perpendicular binding is contributed majorly by the σ bonding orbital that is formed from the HOMO of phenoxides. To sum up, protons bind to all substituted phenoxides through their HOMO in the same way, but CO$_2$ binds to their HOMO or HOMO-1 depending on the sterics of the ortho substituents. For the substituted phenoxides without bulky ortho substituents, the proton and CO$_2$ bindings are both associated to the HOMO, which is the origin of the previously observed LSR between $pK_{a}$ and $pK_{CO2}$. For the substituted phenoxides with bulky ortho substituents, CO$_2$ and the proton bind through different MOs whose energy level is influenced differently by substituents with different σ or π characteristics (inductive or resonance). In addition, due to the negative electrostatic potential (ESP) around the oxygens in CO$_2$ (SI Appendix, Fig. S5A), the phenoxide-CO$_2$ adduct would be less stabilized by the ortho C-H in phenoxide with an ESP of −80 kcal/mol (SI Appendix, Fig. S5B) than by the ortho methyl in C2 with an ESP of −73 kcal/mol (Fig. 5C). The stabilization effect of ortho alkyl groups through a noncovalent interaction (NCI) could be explicitly visualized by the NCI map. As the ortho substituent gets bulkier from -H (SI Appendix, Fig. S6A) to methyl (SI Appendix, Fig. S6B) and then to tert-butyl (SI Appendix, Fig. S6C), the green isosurface representing attractive NCI between the bound CO$_2$ and the ortho substituent becomes larger. As a result of the discussed effects, increasing the bulkiness of ortho substituents on phenoxide can decouple the CO$_2$ and proton binding energies. The system will then be allowed to move beyond the LSR and access the regions with more favorable energetics, namely, higher water tolerance and stronger CO$_2$ binding.

Another advantage of the GA search is the large and diverse dataset it generates that could be utilized to train predictive ML models. As is shown in SI Appendix, Fig. S7, the neural network (NN) model trained on the GA dataset does an excel-

![](https://www.pnas.org/content/119/24/9339.fig)
the electronic structure and dynamics of the molecular fragments would be required. Still, the predictive ML model is cost-wise suited for initial screening for favorable energetics at the semiquantitative level.

**Experimental Validation of Theoretical Trend and Design Principle.** To further validate the trends and design principles obtained from computational screening and GA search, we synthesized six compounds (Fig. 6A) including lithium phenoxide (C1) (28), C2, 2,6-diisopropylphenoxide (C3), catechoxide (C4) (29), 2-nitrophenoxide (C5), and lithium trifluoroethoxide (C6) (30). Their corresponding DFT-calculated pK\textsubscript{a} and pK\textsubscript{CO2} are plotted in Fig. 6B. C2 and C3 are located in the lower right region of the phenoxide cluster, with slight off-scaling relationship behavior due to the steric-induced change in bonding orbital as discussed in the previous section. C4 in dianionic form (pK\textsubscript{a}2) binds both proton and CO\textsubscript{2} more strongly than regular phenoxides. C6 is moved to the upper left along the pK\textsubscript{CO2} / pK\textsubscript{a} scaling relation of alkoxide by introducing trifluoromethyl (EWG), thereby trading part of the CO\textsubscript{2} binding strength for higher water tolerance. The CO\textsubscript{2} capturing ability of the compounds are characterized by absorption capacity, defined as the molar ratio of absorbed CO\textsubscript{2} to the capturing agent, determined using the experimental set-up (SI Appendix, Fig. S8) proposed in ref. 19. For synthesized compounds in this work, there exists a positive relationship between pK\textsubscript{a} and absorption capacity. C2 and C6 have absorption capacities comparable to commercial amines (noncyclic, primary, and secondary), and C4 outperforms all commercial amines. Notably, the time needed to reach the absorption equilibrium is c.a. 1 h for alkoxides and phenoxides, which is significantly faster than that of amines (c.a. 3 h) in previous reports (19, 31). This can be attributed to the direct one-step adsorption pathway for alkoxides and phenoxides, unlike the amines that need to undergo deprotonation first. The facile CO\textsubscript{2} adsorption kinetics adds to the merit of alkoxides and phenoxides for DAC applications.

In Fig. 6D, the absorption capacity is converted to the pK\textsubscript{CO2}, assuming the contribution from carbonate and bicarbonate pathways is minor. A similar pK\textsubscript{CO2} – pK\textsubscript{a} relation as in Fig. 4B can be observed of the synthesized compounds. The trend for the absorbance capacity of amines versus pK\textsubscript{a} is less defined due to the use of pK\textsubscript{a1} instead of pK\textsubscript{a2} (no experimental values available). C4 shifts toward the lower left, away from the scaling relationship, probably due to the strong CO\textsubscript{2} binding in its dianionic form and more binding sites available (two sites per molecule). On the other hand, C3 shows less CO\textsubscript{2} binding than expected. As discussed above, bulky ortho groups were
thought to increase CO2 binding due to forcing the CO2 to bond through different MOs; however, C3 (isobutyl groups) shows less CO2 binding than C2 (methyl groups), indicating a weakness of the implicit solvation method employed throughout the study. Since isopropyl groups are bulkier, the H-bond interaction between the bound CO2 and its solvent environment is weaker for C3 than for C2 (SI Appendix, Fig. S9). In addition, the water-inaccessible regions caused by the sterics of the ortho groups are closer to the bound CO2 in C3 than in C2, thus lowering the solvent accessibility of the bound CO2 more for C3. Taking this into account, the $pK_{\text{CO2}}$ calculated when considering explicit water molecules is lower than the implicit solvation results by 0.9 units for C3. In contrast, little change is observed for C2. These results suggest the overestimation of CO2 binding for phenoxide with highly bulky ortho groups (e.g., iso-propyl and tert-butyl), due to the implicit solvation model giving the incorrect solvation free energy. After correcting the $pK_a$ in Fig. 6D using the explicit solvation results, C3 would shift closer to the scaling relation, while C2 stays about the same. In this work, we focus on presenting the computational workflow for inverse molecular design; hence, compounds with relatively simple synthetic route are chosen mainly for the purpose of validation.

**Conclusions**

In conclusion, we explored amines, alkoxides, and phenoxides with a series of theoretical and computational methods in search of the optimal sorbent for aqueous DAC of CO2. DFT calculations are first performed to study the bonding nature and energetics of autoionization, CO2 binding, and deprotonation on substituted amines. The anionic deprotonated amino group is found to be the species that binds CO2. We discovered a convex hull relationship that prohibits the optimization of both $\Delta G_{\text{bind}}$ (binding strength) and $\Delta G_{\text{auto-ionization}}$ (concentration of the binding species) and an LSR between $pK_{\text{CO2}}$ and $pK_{\text{a}}$ of amines. Alkoxides and phenoxides are then proposed as better sorbents for their improved water tolerance and more realistic solvation free energy into the screening process includes QM/MM Monte–Carlo (or other advanced sampling techniques) (32) and ML models trained on all-QM datasets (33).

Although C2 is still not as good as C4, we note that it has been improved by 76.5% in absorption capacity (from 0.17 to 0.30) compared to the unsubstituted phenoxide C1, and its meta/para sites are still available for further substitution. Meanwhile, the C4 may also be further improved by introducing ortho bulky groups to exploit the steric effect that we are currently pursuing. In this work, we focus on presenting the computational workflow for inverse molecular design; hence, compounds with relatively simple synthetic route are chosen mainly for the purpose of validation.

![Fig. 6.](https://doi.org/10.1073/pnas.2123496119)
favorable capture stoichiometry. All three molecular systems are found to bind CO$_2$ in primarily an $\sigma$-fashion and follow the same pK$_{CO_2}$–pK$_\text{L}$ LSR. Several high-performance alkoxides are proposed from the computational screening. Since there is no opportunity of pK$_{CO_2}$–pK$_\text{L}$ decoupling on phenoxides, we apply GA global optimization combined with SQM calculations to explore the large chemical space of substituted phenoxides. Several promising molecules with off-LSR energetics are discovered from the GA searches and validated by DFT calculations. The top-scoring molecules are then studied by bonding analysis and PES scan to understand the origin of the off-LSR behavior; bulky ortho substituents force the CO$_2$ adduct into a perpendicular configuration that binds through HOMO, while the proton still binds through HOMO-1 in the in-plane configuration, thus decoupling the pK$_{CO_2}$ and pK$_\text{L}$. Several substituted phenoxide and alkoxide molecules are synthesized for CO$_2$ absorption capacity measurement to validate the computational results. The overall pK$_{CO_2}$–pK$_\text{L}$ trend and off-LSR behavior of specific molecules (bulky groups or hydroxyl on ortho positions) are observed in the experiments, which is consistent with the theory. The inverse molecular design workflow presented in this work is highly generalizable and can be readily adapted for optimizing other functional molecules for various applications.

**Methods**

**Computational Methods.** Geometry optimization, vibrational analysis, and implicit solvation model calculations are performed using the Gaussian 16 program (34) (revision C01). The geometry optimizations and PES scans are performed using B3LYP functional (35, 36) with def2-TZVP basis sets (37) and D3 correction (Becke-Johnson damping) (38) to better account for the dispersion interactions. Harmonic vibrational frequencies are computed on each optimized geometry to make sure that all reaction intermediates have no imaginary frequency. The entropic and thermo-statistical contributions ($\delta G^\text{thermo}$) are calculated by the rigid rotor harmonic oscillator approximation on optimized geometries at 298.15 K and 1 atm. The value of $\delta G^\text{thermo}$ is calculated similarly from the free energy change of the CO$_2$ binding process (LFER parameters for pK$_{CO_2}$ are set to $c_0 = 1$ and $c_1 = 0$ due to a lack of available experimental data): 

$$\Delta G^\text{deprot} = G(A^+ + \text{CO}_2) - G(A^+) - G(\text{CO}_2)$$

$$\Delta G^\text{deprot} = c_0 \cdot \frac{\Delta G^\text{deprot}}{\ln(10)^\text{RT}} + c_1.$$ 

According to our test, the DFT protocol of this work outperforms DLPNO-CCSD(T)/def2-QZVPP (45) and quantum chemistry composite method CBS-QB3 (46) in reproducing experimental pK$_\text{L}$ of alkoxides (SI Appendix, Fig. S1).

Molecular orbital analysis, Hirshfeld population analysis, NCI analysis, and ESP mapping are performed using the Multiwfn program on the converged wavefunctions from DFT calculation (47). The configurational sampling under explicit solvation is performed using the genmer module in the Molcul program (48).

In the computational screening section, the substituent pool includes (24 in total) the following: -OH, -CH$_3$, -CN, -CH$_2$-C$_3$H$_6$, -CH$_2$-C$_2$H$_5$, -CH$_2$-CH$_2$, -CH$_2$Cl, -CF$_3$, -CHO, -COCH$_3$, -COOC$_2$H$_5$, -OCOCH$_3$, -F, Cl, Br, -OC$_2$H$_5$, -OCH$_3$, -SO$_2$CH$_3$, -SO$_2$OCH$_3$, -SO$_2$H, -NO$_2$, and -CHNH. Note that the hydroxyl and amino groups are not included in the pool to avoid difficulty in pK$_\text{L}$ determination of polyamines and alkanolamines. Tent butyl is also not included because its large volume and rigidity can cause numerical instabilities due to atom overlapping during the SMILE-to-XYZ conversion and conformational search by the MMFF94 force field.

Each molecule is represented by a one-dimensional (1D) vector with five (number of substitution sites) elements, each representing a substituent. The representation (noted as gene representation of molecules) can reversibly interconvert into or from a SMILES representation (49). The SMILES string is converted to XYZ coordinate using Open Babel package (50), and sufficient stochastic conformational search is performed at the MMFF94 level to obtain the most stable conformation (51).

Connectivity is checked after geometry optimization to make sure there is no unexpected bond dissociation that suggests instability of the molecule under certain charge states. A data point is discarded if any of the involved species (neutral, deprotonated, CO$_2$ adduct) is found unstable.

SQM calculations are performed using the xTB package (52) for the geometry optimization and energy evaluation throughout the GA search. GFN1-xTB tight binding method with GBSA model for describing implicit solvation by water is employed in the high-throughput computational screening section for its low computational cost and comparable accuracy to DFT methods in terms of geometry and thermochemistry (53). The pK$_\text{L}$ and pK$_{CO_2}$ values are calculated following the fitting and LFER procedure as described in the DFT section.

The GA search is performed using an adapted version of the molGA code (15). The population size, mutation rate, and convergence criterion are set to 100 candidates, 33%, and 100 generations, respectively. The search goal is to minimize the $\Delta G^\text{deprot}$ while keeping the $\Delta G^\text{deprot}$ higher than that of OH$^-$. Ten independent GA searches are performed, and the top-scoring candidates are collected from the final populations. DFT calculations are performed on those candidates thereafter to obtain more accurate energetics.

The ML model for the fast prediction of molecular properties is a multilayer perceptron NN with four rectified linear units and two hidden linear layers, implemented using pytorch library (54). Each phenoxide molecule is converted into a 1D vector with 24 (number of substitution sites) × 5 (number of substitution sites) = 120 binary elements via positional one-hot encoding (SI Appendix, Scheme S1). The NN is trained on the dataset of sampled substituted phenoxides labeled with SQM-calculated properties, with 80% of the data as a training set. 10-fold cross validation is performed, and the best NN is selected that yields the least mean square error (MSE) on the cross-validation set. The NN is evaluated on the full training set and the test set (20% of the data).

Experimental Methods. Synthesis and manipulation of compounds were carried out in open air unless otherwise mentioned. For air- and moisture-sensitive procedures, manipulations were carried out in a glovebox or using standard Schlenk techniques under inert atmosphere of nitrogen. All solvents and reagents were purchased from commercial vendors and used without further purification unless otherwise noted. Pentane and toluene used an inert atmosphere syntheses and/or manipulations were degassed by sparging with argon.
and dried by passing through columns of neutral alumina or molecular sieves. Water used during inert atmosphere synthesis and/or manipulations was degassed using an active vacuum for several hours. All deuterated solvents were purchased from Cambridge Isotope Laboratories, Inc. Deuterated methanol, DMSO, benzene, and water were degassed and methanol and DMSO stored over activated 3 Å molecular sieves prior to use. The compounds in Fig. 6A are synthesized using two routes that are described as follows.

**Synthesis using LiOH.** This synthetic route was used for deprotonation of tri-fluoroethanol, hexafluoropropanol, 2-nitrophenol, and 2,6-dimethylphenol. Under an N2 atmosphere, 25 mmol of alcohol was combined with 25 mmol of lithium hydroxide in 20 ml dry methanol. The reaction was refluxed overnight and then dried under vacuum to give a solid.

**Synthesis using n-butyl lithium.** This synthetic route was used to deprotonate phenol, 2,6-disopropylphenol, and catechol. Alcohols were first stirred with toluene for an hour and then dried under vacuum to remove excess water. Under an N2 atmosphere with 50 ml of dried pentane from the solvent system and 25 mmol of dried alcohol, 25 mmol (or 50 mmol for catechol) 1.6 M n-butyl lithium in hexane in an ice bath was dropped addwise at −78 °C. The mixture was stirred at −78 °C for 1 h and then stirred for an additional 24 h at room temperature. The solvent was removed under reduced pressure to leave a colorless solid.

NMR spectroscopy was used to confirm the identity and purity of the synthesized compounds. 1H NMR spectroscopy was performed on a 500 MHz Bruker Avance G500 with a 5 mm probe or on a 500 MHz Bruker DRX 500 spectrometer with a TCI cryoprobe. 13C NMR spectra were recorded on a 500 MHz Bruker DRX 500 fitted with a TCI cryoprobe. All NMR spectra were acquired at room temperature and referenced to residual 1H or 13C resonances of the deuterated solvent (D, CD3OD, 6.33; D2O, 6.79; DMSO-D6, 6.25; CDCl3, 7.26; DMSO-D6, 7.16). 1H NMR (CD3OD, 6.49; D2O, 6.8; DMSO-D6, 4.93).

**Lithium phenoxide (C7).** 1H NMR (500 MHz, D2O) δ 6.82 (t, J = 7.0 Hz, 2H), 6.37 (s, 1H), 6.10 (s, 1H); 13C NMR (126 MHz, D2O) 61.76, 128.88, 120.03, 109.97.

**Lithium dimethylphenoxide (C8).** 1H NMR (500 MHz, CD3OD) δ 7.04 (d, J = 6.9 Hz, 2H), 6.71 (t, J = 7.3 Hz, 1H), 2.15 (s, 6H); 13C NMR (126 MHz, CD3OD) δ 63.16, 163.46, 128.21, 127.04, 113.61, 17.34.

**Lithium 2,6-diisopropylphenoxide (C3).** 1H NMR (500 MHz, D2O) δ 7.03 (d, J = 7.4 Hz, 2H), 6.60 (t, J = 7.4 Hz, 1H), 3.37 (q, J = 6.9 Hz, 2H), 1.13 (d, J = 6.9 Hz, 12 H); 13C NMR (126 MHz, D2O) δ 160.14, 138.31, 122.81, 114.19, 25.82, 23.00.

**Lithium catecholate (C4).** 1H NMR (500 MHz, D2O) δ 6.27 (s, 2H), 5.94 (s, 2H); 13C NMR (126 MHz, D2O) δ 161.56, 116.33, 112.74.

**Lithium 2-nitrophenoxide (C5).** 1H NMR (500 MHz, D2O) δ 7.89 (d, J = 8.4 Hz, 1H), 7.36 (t, J = 7.7 Hz, 1H), 6.79 (d, J = 8.7 Hz, 1H), 6.51 (t, J = 7.6 Hz, 1H); 13C NMR (126 MHz, D2O) δ 165.77, 137.50, 136.06, 126.42, 125.43, 113.06.

**Lithium 2,2-trifluoroethoxide (C6).** 1H NMR (500 MHz, CD3OD) δ 3.86 (q, J = 9.2 Hz, 2H); 13C NMR (126 MHz, CD3OD) δ 129.35, 127.19, 124.93, 122.75, 61.83, 61.56, 61.30, 61.06; 19F NMR (565 MHz, D2O) δ -76.44.

The CO2 absorption capacity measurements were made following the procedure previously described in ref. 19 by Puxty et al.; a total of 10 ml of 0.5 M alkoxide solution was placed in a weighed 20 ml vial with septum screw top and stir bar (SI Appendix, Fig. S84). First, the mass change due to evaporation was recorded by placing the vial in a 40 °C bath and sparging with N2. The inlet needle was never placed directly into solution, only the headspace. The change in mass was measured 8 times over a period of 20 min. Next, the gas inlet was changed to 10% CO2. The change in mass was measured every minute for the first 10 min, and then every 5 min for 1 h. To determine the overall CO2 absorption, the mass change due to evaporation alone was subtracted from mass change when 10% CO2 was used to give the total mass gained due to CO2 absorption (SI Appendix, Fig. S88). This method was validated against the original data from Puxty et al. (19) using ethylenediamine as the standard.

### Associated Content

**Supporting Information.** The following files are available free of charge: schematics of the conversion between data representations; IFER fitting plots of DFT and SOM calculations; scatter plots showing correlation among pKCO2, pK1, and pK2 of amines; the evolution diagram of the ΔGabs during the GA searches; the ESP and NCI map of the investigated substituted phenoxides and CO2 adducts; validation of the trained NN model; experimental setup for CO2 absorption capacity measurement; and the H-bond geometry and bond order between water and CO2-bound phenoxides.

**Data Availability.** All study data are included in the article and/or SI Appendix.

**ACKNOWLEDGMENTS.** This research was made possible thanks to generous support from the Alfred P. Sloan Foundation. The computations presented in this study were performed using the Extreme Science and Engineering Discovery Environment (XSEDE) supported by the National Science Foundation. Z.Z. thanks Daniel Bim for helpful discussions.
33. Y. Basdogan et al., Machine learning-guided approach for studying solvation environments. J. Chem. Theory Comput. 16, 633–642 (2020).
34. M. J. Frisch et al., Gaussian 16 (2016).
35. A. D. Becke, Becke’s three parameter hybrid method using the LYP correlation functional. J. Chem. Phys. 98, 5648–5652 (1993).
36. J. Tirado-Rives, W. L. Jorgensen, Performance of B3LYP density functional methods for a large set of organic molecules. J. Chem. Theory Comput. 4, 297–306 (2008).
37. F. Wengend, R. Ablrichs, Balanced basis sets of split valence, triple zeta valence and quadruple zeta valence quality for H to Rn: Design and assessment of accuracy. Phys. Chem. Chem. Phys. 7, 3297–3305 (2005).
38. S. Grimme, S. Ehrlich, L. Goerigk, Effect of the damping function in dispersion corrected density functional theory. J. Comput. Chem. 32, 1456–1465 (2011).
39. A. V. Marenich, C. J. Cramer, D. G. Truhlar, Universal solvation model based on solute electron density and on a continuum model of the solvent defined by the bulk dielectric constant and atomic surface tensions. J. Phys. Chem. B 113, 6378–6396 (2009).
40. Y. Zhao, N. E. Schultz, D. G. Truhlar, Design of density functionals by combining the method of constraint satisfaction with parameterization for thermochemistry, thermochemical kinetics, and noncovalent interactions. J. Chem. Theory Comput. 2, 364–382 (2006).
41. J. Ho, A. Klamt, M. L. Coote, Comment on the correct use of continuum solvent models. J. Phys. Chem. A 114, T3442–T3444 (2010).
42. E. Paulechka, A. Kazakov, Efficient DLPNO-CCSD(T)-based estimation of formation enthalpies for C-, H-, O-, and N-containing closed-shell compounds validated against critically evaluated experimental data. J. Phys. Chem. A 121, 4379–4387 (2017).
43. F. Neese, F. Wennmohs, U. Becker, C. Riplinger, The ORCA quantum chemistry program package. J. Chem. Phys. 152, 224108 (2020).
44. K. S. Alongi, G. C. Shields, “Theoretical calculations of acid dissociation constants: A review article” in Annual Reports in Computational Chemistry, (Elsevier, 2010), pp. 113–138.
45. E. Paulechka, A. Kazakov, Efficient DLPNO-CCSD(T)-based estimation of formation enthalpies for C-, H-, O-, and N-containing closed-shell compounds validated against critically evaluated experimental data. J. Phys. Chem. A 121, 4379–4387 (2017).
46. G. P. F. Wood et al., A restricted-open-shell complete-basis-set model chemistry. J. Chem. Phys. 125, 094106 (2006).
47. T. Lu, F. Chen, Multifcn: A multifunctional wavefunction analyzer. J. Comput. Chem. 33, 580–592 (2012).
48. K. S. Alongi, G. C. Shields, “Theoretical calculations of acid dissociation constants: A review article” in Annual Reports in Computational Chemistry, (Elsevier, 2010), pp. 113–138.
49. D. Weininger, SMILES, a chemical language and information system. 1. Introduction to methodology and encoding rules. J. Chem. Inf. Comput. Sci. 28, 31–36 (1988).
50. N. M. O’Boyle et al., Open Babel: an open chemical toolbox. J. Cheminform. 3, 33 (2011).
51. N. Yoshikawa, G. R. Hutchinson, Fast, efficient fragment-based coordinate generation for Open Babel. J. Cheminform. 11, 49 (2019).
52. S. Grimme, XTB, A tight-binding Quantum Chemistry Code for Large Molecules (University of Bonn, 2019).
53. C. Bannwarth, S. Ehler, S. Grimme, GFN2-xTB: An accurate and broadly parametrized self-consistent tight-binding quantum chemical method with multipole electrostatics and density-dependent dispersion contributions. J. Chem. Theory Comput. 15, 1652-1671 (2019).
54. A. Paszke et al., Pytorch: An imperative style, high-performance deep learning library. Adv. Neural Inf. Process. Spot. 32, 8026–8037 (2019).