nature chemistry

ARTICLES

PUBLISHED ONLINE: 2 NOVEMBER 2014 | DOI: 10.1038/NCHEM.2099

Discovering chemistry with an *ab initio* nanoreactor

Lee-Ping Wang, Alexey Titov†, Robert McGibbon, Fang Liu, Vijay S. Pande and Todd J. Martínez*

Chemical understanding is driven by the experimental discovery of new compounds and reactivity, and is supported by theory and computation that provide detailed physical insight. Although theoretical and computational studies have generally focused on specific processes or mechanistic hypotheses, recent methodological and computational advances harken the advent of their principal role in discovery. Here we report the development and application of the *ab initio* nanoreactor—a highly accelerated first-principles molecular dynamics simulation of chemical reactions that discovers new molecules and mechanisms without preordained reaction coordinates or elementary steps. Using the nanoreactor, we show new pathways for glycine synthesis from primitive compounds proposed to exist on the early Earth, which provide new insight into the classic Urey–Miller experiment. These results highlight the emergence of theoretical and computational chemistry as a tool for discovery, in addition to its traditional role of interpreting experimental findings.

Experimental chemistry often plays the principal role in discovering new compounds and proposing new reaction mechanisms, and computational chemistry provides valuable support by arbitrating between competing proposed mechanisms. Recent algorithmic and computational advances, including those that leverage graphics processing unit (GPU) architectures[1–4], could open the door to using computation not only to arbitrate different hypotheses, but also as a discovery tool to reveal new fundamental chemical mechanisms. Our experimentally inspired[5] *ab initio* nanoreactor accomplishes this using an *ab initio* molecular dynamics (AIMD) simulation of freely reacting molecules, coupled with automatic analysis and refinement methods to build a quantitatively accurate reaction network. By seeding the nanoreactor with diverse reactants available in various environments, such as the early Earth or the upper atmosphere, we explore reactivity and discover new reaction schemes. This approach will help guide experiment by posing new hypotheses and suggesting novel experiments.

The statistical rarity of activated chemical reactions restricts most AIMD studies to specific transformations along a chosen reaction coordinate or collective variable[6–8]. A promising approach to overcome the rarity of reactive events has been the application of predefined heuristic rules[9–11] or geometric searching[12,13] to generate new molecules and reaction networks. In contrast, the nanoreactor discovers molecules and reactions based only on the fundamental equations of quantum and classical mechanics. Reactions occur freely without preordained reaction coordinates or elementary steps.

Although recent advances in AIMD provide much computational relief, these simulations nevertheless remain costly for sampling large numbers of reactive events. We overcome this difficulty by incorporating new acceleration techniques in the nanoreactor. A virtual piston enhances reactivity by periodically pushing molecules towards the centre of the nanoreactor, which greatly increases the frequency of collisions and barrier crossings (see Supplementary Fig. 1). This evokes ideas from high-pressure and shock-wave simulations[14–16], with the key difference that the periodic forcing increases the number of barrier crossings through ballistic collisions rather than inducing an equilibrium high-pressure regime. Furthermore, we use an approximate Hartree–Fock (HF) ansatz to access large simulation sizes (hundreds of atoms) and long timescales (hundreds of picoseconds). Sampling of chemical space at this approximate level is augmented by subsequent energy refinement of the discovered reaction pathways using more-quantitative methods such as density functional theory (DFT). This strategy exploits the fact that the qualitative topography of the energy landscape is described well by methods that may not provide quantitative estimates of reaction rates. For example, HF is well-known to predict chemically reasonable molecular structures[17], even though DFT[18] and more-sophisticated wavefunction methods[19] are more accurate for thermochemistry and barrier heights.

The nanoreactor achieves its goal of broadly exploring reaction pathways by taking an intermediate stance between physically realistic simulation and rule-based enumeration approaches. The *ab initio* simulation ensures that reaction trajectories obey physical equations of motion and avoids a combinatorial explosion of possibilities, and the occurrence of reactions is accelerated by explicitly not aiming to replicate the physicochemical conditions of any one environment. The pathways that result from energy refinement are applicable to any thermodynamic setting by providing reaction parameters (for example, concentration, temperature) as input variables to a kinetic model. This approach is valid as long as the relevant reactions are sampled at least once and included in the knowledge base. To ensure complete sampling can be difficult and it would be premature to claim that we have achieved this for the prototypical cases presented in this paper. Here we focus on introducing the nanoreactor, present some newly discovered pathways from nanoreactor simulations and discuss the broader implications of discovery-based theoretical methods.

Results and discussion

Insight into the synthesis of a diverse set of products. We discuss two nanoreactor simulations on contrasting systems. The first starts with a homogeneous collection of acetylene molecules, which we chose because of the well-known tendency of acetylene to polymerize into larger molecules. The second one is an idealization of the classic 'Urey–Miller' experiment[20], and includes several compounds postulated to exist in the early Earth

Department of Chemistry, Stanford University, Stanford, California 94305, USA; †Present address: Advanced Micro Devices, Sunnyvale, California 94088, USA. *e-mail: toddjmartinez@gmail.com

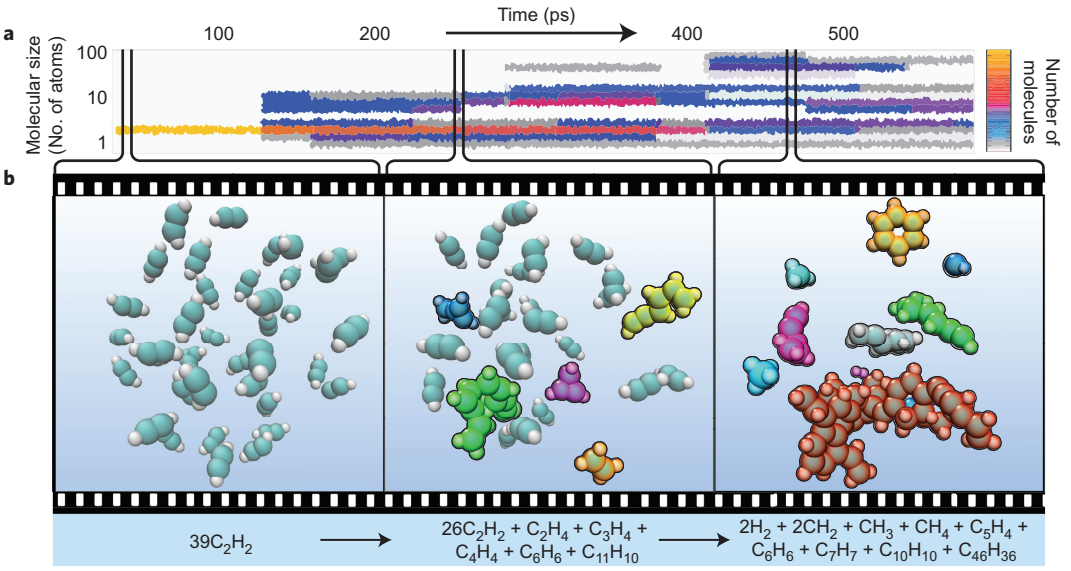

Figure 1 | Timeline of a nanoreactor simulation trajectory (see Supplementary Video 1 for a movie clip). a, Molecular size distribution as a function of simulation time. **b**, Left: Simulation begins with a collection of acetylene molecules (C, teal; H, white). New molecules are automatically highlighted with molecule-specific colours to indicate the observed reactivity. Middle: Simple products appear first, including short polymeric species (green, yellow) as well as ethylene (orange) and cyclopropene (violet). Right: At longer simulation times the molecular size distribution becomes considerably wider and more than half the atoms form a large molecule that contains multiple aromatic rings (red). A long-lived, inert benzene molecule is also formed (gold).

atmosphere (hydrogen, ammonia, methane, carbon monoxide and water). The Urey–Miller simulation differed from the experimental conditions in that the virtual piston was used in place of electric sparks, although both methods provided an energy input to accelerate barrier crossings. Both simulations consisted of many initial reactant molecules (50–100) to sample a large reaction space.

Figure 1 illustrates the acetylene nanoreactor simulation (see the movie clip in Supplementary Video 1). Molecules freely reacted with each other over the course of the simulation. The piston accelerated the reaction rate by oscillation with a period of 2 ps (4,000 time steps). Nearly 100 distinct products were formed after ~500 ps simulation time (one million time steps), including methane, ethylene, cyclopropene, benzene and larger polymeric species with both aliphatic and aromatic character (Supplementary Fig. 3). We visualized the simulation trajectory using a machine-learning algorithm to identify new products and automatically highlighted them in molecule-specific colours.

The diversity of the discovered compounds is surprisingly rich. Previous experiments on acetylene reactivity at high pressure[21,22] indicate an increase in the number of single and double C–C bonds and a decrease in the number of triple bonds; a combination of linear and branched conjugated chains formed rather than a covalently bonded single crystal. The nanoreactor produced some linear and branched conjugated chains similar to those in the experiment, but there were also many new motifs, including aromatic rings, allenes and a smaller number of antiaromatic and highly strained rings. As the goal of the nanoreactor is to discover new reactivity independently of specific experimental conditions, it is encouraging that we not only reproduced some of the observed chemistry from the high-pressure experiments, but also found a greater diversity of chemical species, which may be important in other settings. This diversity is, in part, because of the high kinetic energy imparted by the piston, which corresponded to instantaneous temperatures as high as ~10,000 K; at such temperatures, electronic excitations may be thermally accessible. Although the resulting multistate non-adiabatic dynamical effects could be included[23], the nanoreactor currently ignores them, consistent with its primary goal to sample reaction space rather than realistically model a particular physical process.

The Urey–Miller-inspired simulation generated a starkly different collection of molecules, with much smaller products. Among the discovered products were the natural amino acid glycine, the unnatural amino acids α-hydroxyglycine and α-aminoglycine, and a reduced analogue of alanine in which a geminal diol replaced the carboxyl group (see Supplementary Fig. 4). Additional products discovered included urea, ethylene glycol and isocyanic acid, all of which have also been detected in meteorites that may have delivered organic molecules to the early Earth[24]. A few illustrative examples of the discovered reactions are provided in Supplementary Figs 5–9. These examples include reactions catalysed by surrounding ammonia or water molecules that act as proton shuttles.

A complex web of reaction pathways. In addition to the high diversity of products, the nanoreactor simulation also offers insight into how the products were formed. The molecular dynamics pathway that connects stable reactant and product species was used to locate a corresponding minimum-energy path (MEP). Using these MEPs, we built a network of reaction mechanisms that linked products with reactants. More than 700 distinct reactions were found in the Urey–Miller simulation, with a wide distribution of reaction energies and barrier heights (see Supplementary Fig. 2). A significant fraction of the reactions occurred with barriers <50 kcal mol^{-1}, which indicates they may be kinetically viable under ambient conditions.

To derive chemical insight from a complex web of reactions can be challenging. If we are interested mainly in a particular compound, we can map out the local network of closely related compounds (namely, the molecules that appear on either side of chemical equations that lead to the compound of interest). To do this, we focus on a particular molecule in the reaction network and investigate the energetics of the reactions it is involved in. Figure 2 shows one such representation of a reaction network derived from the Urey–Miller nanoreactor (three-dimensional (3D) view in Supplementary Video 2), which includes hundreds of products. Here we focused on a particular molecule (urea, red sphere) and visualized the reactions it was involved in, which leads to a second tier of molecules (blue spheres). The coloured arrows indicate chemical reactions and arrowheads indicate one

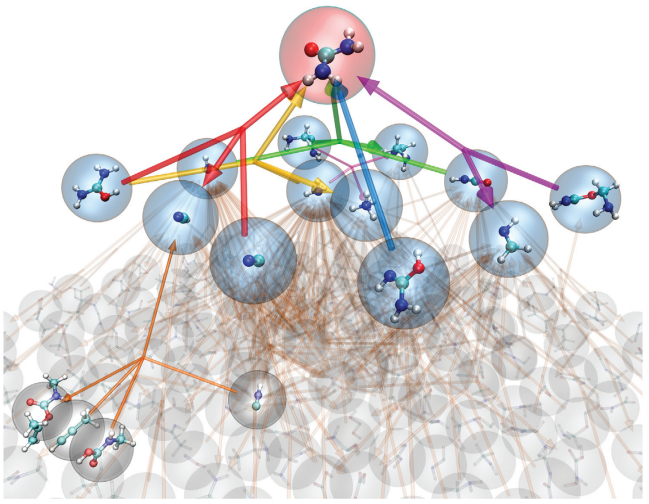

Figure 2 | Pyramid representation of reaction network focused on a product molecule of interest (3D view in Supplementary Video 2). The initial reactants were H_2, H_2O, NH_3, CH_4 and CO. Compounds (C, teal; H, white; N, blue; O, red) are shown in spheres, and reactions (that is, chemical equations) are indicated using coloured arrows. Arrowheads indicate one side of the chemical equation, although reactions can occur in either direction. The chosen molecule (urea) is highlighted in red, and molecules directly involved in reactions with urea are highlighted in blue. Reactions more than one step removed from urea are mostly blurred out to show the high connectivity and complexity in the overall graph, with a single reaction highlighted (grey spheres, bottom left).

side of the chemical equation, although reactions can occur in either direction. As each molecule is involved in reactions with so many others, the third tier of molecules (grey spheres) numbers in the hundreds and cannot be represented clearly. In the foreground, carbamimidic acid, $H_2NC(NH)OH$, tautomerizes to urea, $CO(NH_2)_2$, via proton transfer (blue arrows). Formaldimine, H_2CNH, also reacts with urea to form an ester adduct (violet arrows, right). Many of these molecules are found in interstellar clouds, and the pathways outlined here may be instructive for reactions that happen in a variety of environments, including interstellar space[25,26].

Following a specific reaction. Focusing on a specific molecule allows us to trace the synthetic pathways that lead from the starting materials. Figure 3 shows such a collection of pathways leading to glycine. Here glycine was formed through several distinct pathways that involved reaction barriers of less than 40 kcal mol^{-1}. Formaldimine (Fig. 3, centre) is a key intermediate that participates in three of the four pathways. In one pathway, formaldimine combines with H_2O and CO in a termolecular reaction, and in the other two pathways it combines with formic acid (HCOOH) and proceeds through a singlet carbene intermediate. Aminomethanol (H_2NCOH, Fig. 3, right) is another key intermediate—it is a precursor to formaldimine, but it can also react with CO directly to yield glycine.

Formaldimine, formaldehyde and formic acid are among the most highly connected compounds in the reaction network, participating in more than 40 reactions with other species (the other two species with such a high connectivity are methanol and hydrogen cyanide, along with the initial reactants). These highly connected compounds have in common the ability to react via several different types of pathways; for example, formic acid is found to participate in proton transfer, nucleophilic addition and dehydration reactions. Formic acid is formed easily from the starting materials by the addition of water to carbon monoxide, whereas formaldimine requires several more elementary steps because of the need to form a C=N double bond. The C=N double bond of formaldimine participates in many addition reactions as either the nucleophile or electrophile, and leads to a diverse collection of primary amines and secondary imines. The glycine synthesis pathways involve H_2 only once in the hydrogenation of formic acid to yield methanediol, and CH_4 never appears (it is highly inert). This supports previous proposals that biomolecules may have formed with little participation from these highly reducing compounds[24].

Emergence of higher-order chemical principles. Higher-order chemical principles emerge naturally from simulation and analysis in the nanoreactor. For example, the acetylene nanoreactor formed a large number of C–C bonds, whereas the Urey–Miller nanoreactor did not. Many alternative pathways that compete with C–C bond formation are available in the Urey–Miller system, most notably carbon–heteroatom bond formation via nucleophilic addition, which involves a lower activation energy.

Another interesting observation is that the acetylene simulation forms very large molecules, which include a single large species that comprises more than 70 atoms, whereas the Urey–Miller simulation forms much smaller molecules (up to 16 atoms). This is because the sum total of bond orders across the entire simulation is roughly conserved, a natural consequence of electron conservation. The acetylene nanoreactor starts with a large number of triple bonds that can be traded to make more single bonds between molecules, whereas most of the Urey–Miller reactants are fully saturated molecules. Without double and triple bonds, a bimolecular reaction of two molecules that yields a larger product must also eliminate a smaller product, which leads to a quasi-equilibrium in the molecular size distribution.

The essential catalytic role of water and ammonia illustrates the significance of the solvent in reducing the barrier of important pathways in which hydrogen atoms or protons are transferred. In Fig. 3, more than half the elementary steps involve one or two catalytic water and/or ammonia molecules that participate by acting as a proton wire. For example, the barrier to the dehydration of methanediol (to yield formaldehyde) is lowered by more than 15 kcal mol^{-1} (from 43.5 to 28.3 kcal mol^{-1}) by the presence of a catalytic water molecule. In the three elementary steps by which

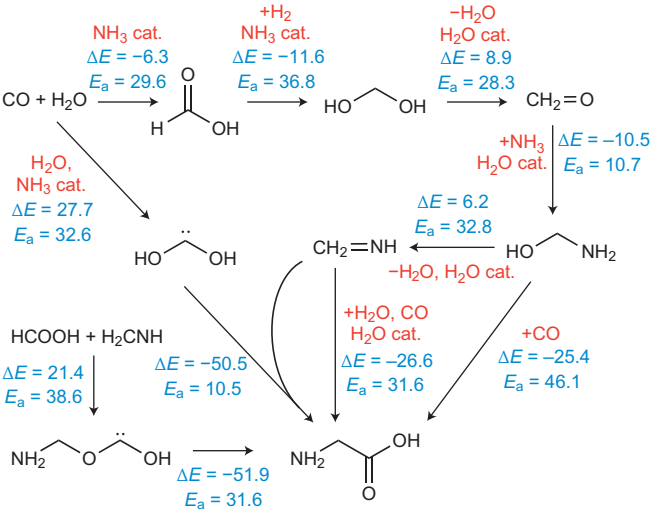

Figure 3 | Sequence of elementary reaction steps derived from the nanoreactor simulation. The sequence begins with the fundamental reactants (CO, H_2, H_2O and NH_3) and ends with the amino acid glycine. Glycine (bottom centre) is formed via four different pathways, three of which involve formaldimine (centre) and two of which involve singlet carbene intermediates. Reaction energies (ΔE) and activation barriers (E_a) calculated using DFT are provided in kcal mol^{-1}. Molecules labelled 'cat.' participate catalytically as proton shuttles.

carbon monoxide is hydrogenated to yield formaldehyde (Fig. 3, top) a water molecule is incorporated temporarily and the highest barrier is 36.8 kcal mol^{-1}; the direct hydrogenation is much less favourable as it has a barrier of 69.5 kcal mol^{-1}. In an aqueous environment, the presence of many solvent molecules would further facilitate such chemistry by stabilizing highly polar or temporarily charged species (for example, H_3O^+ or NH_4^+). Thus hydrogen-bonding solvents, such as water, play both implicit and explicit roles; we plan to include implicit solvent effects to improve the accuracy of the energy refinement for condensed-phase conditions.

The future of the nanoreactor approach. The provided examples demonstrate that: (1) the *ab initio* nanoreactor not only finds many reactions that are well-known from experimental chemistry, but also discovers new pathways not characterized previously; (2) many of these reactions proceed through low-to-moderate reaction barriers, in spite of the simulation conditions; (3) finally, some of the reactivity is complex and highly concerted, and is thus unlikely to be discovered through heuristic rule-based approaches. The termolecular reaction to yield glycine in Fig. 3 and the acid-catalysed ring-opening in Supplementary Fig. 9 are examples of complex mechanisms because three bonds are broken and two bonds are formed in a single barrier crossing. The nanoreactor discovers many termolecular reactions because of the way it accelerates molecular collisions; although these reactions are rare in the gas phase, they can become relevant when two or more reactants form a preassociated complex[27].

Here we show two nanoreactor simulations with dramatically different results: the acetylene simulation underwent massive polymerization, whereas the Urey–Miller simulation generated a complex network of reactions, which included several pathways to glycine that pass through formaldimine, formic acid and aminomethanol as intermediates. Many of the discovered reactions are complex and concerted, which highlights the unique utility of the nanoreactor as a purely discovery-based method to generate chemically interesting elementary steps and supplement existing methods reliant on hypotheses and prior expectations. More-recent studies in prebiotic chemistry argue that the early Earth atmosphere was probably much less reducing, containing N_2, CO_2 and possibly even some O_2 (refs 28,29), and thus the prebiotic significance of this study should be taken in the context of the original Urey–Miller experiment rather than the more modern hypotheses of the ancient Earth's atmospheric composition. We anticipate that the nanoreactor will contribute to our future understanding of complex reactivity in natural systems by providing novel hypotheses for reaction pathways and elementary steps in arenas as diverse as catalysis, prebiotic chemistry and astrochemistry.

Methods

The nanoreactor AIMD simulations were performed with the TeraChem quantum chemistry and AIMD software packages[1,2,30–34], using the HF electronic wavefunction and a 3-21G Gaussian basis set to calculate the Born–Oppenheimer potential energy surface. The acetylene simulations used unrestricted HF and employed level shifting[35] to allow for open-shell states, whereas the Urey–Miller simulation used restricted HF. The acetylene simulation used a single initial configuration, whereas the Urey–Miller simulation used four different initial configurations with the same molecules. The equations of motion were integrated numerically using Langevin dynamics with an equilibrium temperature of 2,000 K (also the starting temperature) and a friction coefficient of 7 ps^{-1}. The temperature corresponds to an average kinetic energy of 4.0 kcal mol^{-1} per degree of freedom; the thermal motion rapidly breaks apart non-covalent interactions without breaking the covalent bonds. The calculations were feasible because of the efficiency of TeraChem, which dramatically accelerates the calculation of the Fock operator (especially the Coulomb and exchange operators) by evaluating the two-electron integrals on the GPU. The self-consistent field calculation at each AIMD step was made more robust by using the ADIIS (augmented direct inversion in the iterative subspace) algorithm[36] as a back-up for cases in which the default DIIS algorithm[37] failed to converge. A total of 560 and 1,296 ps time evolution was followed for the acetylene (156 atoms) and Urey–Miller (228 atoms) simulations, respectively. The total computational cost of these calculations was 41,700 (acetylene) and 132,400 (Urey–Miller) central processing units (CPU) and GPU hours; TeraChem uses one CPU core per GPU.

The molecules were restrained to move inside a spherical volume by a boundary potential, with a time-dependent component to increase the occurrence of reaction events:

$$V(r,\,t) = f(t)U\left(r,r_1,k_1\right) + \left(1 - f(t)\right)U\left(r,r_2,k_2\right)$$

$$U(r,r_0,k) = \frac{mk}{2}(r - r_0)^2\theta(r - r_0);\; f(t) = \theta\left(\left\lfloor \frac{t}{T} \right\rfloor - \frac{t}{T} + \frac{\tau}{T}\right)$$

where $k_1 = 1.0$ kcal mol^{-1} Å^{-2}, $r_1 = 14.0$ Å, $k_2 = 0.5$ kcal mol^{-1} Å^{-2}, $r_2 = 8.0$ Å, $\tau = 1.5$ ps, $T = 2.0$ ps, $\lfloor\ \rfloor$ is the floor function and θ is the Heaviside step function. The function $f(t)$ is a rectangular wave that oscillates between one (duration τ) and zero (duration T–τ), and $U(r,\,r_0,\,k)$ is a radial potential that is zero inside the prescribed radius r_0 and harmonic outside. The force constant is multiplied by the atomic mass (in AMU) such that all the atoms at the same radial coordinate are subject to equal acceleration. The rectangular waveform switches the restraint potential between $U(r,r_1,k_1)$ and $U(r,r_2,k_2)$, which forces the atoms with a radial position $8.0 < r < 14.0$ Å towards the centre and causes them to collide. When the sphere is expanded again, the molecules in the simulation diffuse rapidly (because of the high temperature) to fill the larger volume. The rectangular waveform spans a broad frequency range, and thus the applied energy does not preferentially drive any specific mode in the system.

The simulation analysis was performed using graph-theoretical and machine-learning routines in the NETWORKX[38] and SCIKIT-LEARN[39] Python modules. The atomic connectivity for each frame in the nanoreactor AIMD trajectory was determined using covalent radii, and graphs that represent individual molecules were constructed from the connectivity matrix. We identified chemical reactivity in the nanoreactor simulation by searching for changes in the connectivity graphs (that is, molecules) as a function of time. A major challenge in this procedure is the transient appearance of spurious connectivity graphs caused by high-frequency bond vibrations and close contacts during molecular collisions. We addressed this problem by applying a two-state hidden Markov model (HMM) to each time series, in which the observed time series of a given connectivity graph was modelled using an underlying lower-frequency signal:

$$P(Y) = \sum_{X=0,1} P(Y|X)P(X); P(Y|X) = \begin{cases} 0.6, Y = X \\ 0.4, Y \neq X \end{cases};$$

$$X_{i+1} = \begin{pmatrix} 0.999 & 0.001 \\ 0.001 & 0.999 \end{pmatrix} X_i$$

where Y is the observed time series and X is the underlying lower-frequency signal described by a Markov process. The HMM is parameterized by: (1) the probability of correctly observing the hidden signal (60% of the time) and (2) the transition probability matrix for the Markov process (0.1% per time step). The HMMs allowed the algorithm to recognize molecules despite transient disruptions of their connectivity graphs. A reaction in the nanoreactor trajectory is recognized as a sequence of frames in which a set of complete connectivity graphs transforms into a different complete set. The atoms involved in the reaction are extracted from the trajectory, which includes the reactant and product, as well as certain catalytic species that chemically participate but do not change their compositions (for example, a catalytic water molecule in proton transfer). The reactive trajectory segments are used to perform subsequent energy refinements via an MEP search.

To determine accurately the thermochemistry and barrier heights (which can be used to infer reaction rates), the MEP search is performed using more-accurate (and more computationally expensive) electronic structure methods; the increased cost is largely mitigated by the much smaller size of these calculations as they only include the atoms that participate in an individual reaction. We chose to use the B3LYP three-parameter density functional approximation and the larger 6-31+ G(d,p) basis set for its ability to reproduce experimental heats of formation and activation energies in organic chemistry[40,41], but even more accurate and computationally expensive methods, such as coupled cluster[42–44], could also be used. Importantly, the reactive AIMD trajectory segments used to initiate the MEP search contain numerous large-amplitude and high-frequency motions that are orthogonal to the reaction coordinate. Therefore, we carried out the path refinement in several stages, which we briefly summarize here and will cover in detail in a forthcoming publication.

First, the AIMD path end points are energy minimized to obtain optimized reactant and product structures; the sequences of optimization coordinates are joined with the AIMD segment to create a continuous path that connects minimized reactants and products. Next, the path is smoothed with an interpolation algorithm in internal coordinates, which ensures a smooth connecting path that avoids unphysical structures (for example, atoms passing through each other). The interpolated path is used as an initial guess to the string method[45], which provides an estimate of the transition state. From here, the transition state is located using a partitioned rational function optimization algorithm, followed by an intrinsic reaction coordinate (IRC) calculation to reconnect the transition state with the reactant and product. For cases in which the IRC calculation results in different molecules from the initial reactant and product, the IRC-derived end points are used

in the reaction network. The Q-CHEM quantum chemistry software package[46] and the Work Queue distributed computing software[47] were used in the refinement calculations.

Received 16 April 2014; accepted 25 September 2014; published online 2 November 2014

References

1. Ufimtsev, I. S. & Martinez, T. J. Quantum chemistry on graphical processing units. 3. Analytical energy gradients, geometry optimization, and first principles molecular dynamics. *J. Chem. Theory Comput.* **5,** 2619–2628 (2009).
2. Ufimtsev, I. S., Luehr, N. & Martinez, T. J. Charge transfer and polarization in solvated proteins from *ab initio* molecular dynamics. *J. Phys. Chem. Lett.* **2,** 1789–1793 (2011).
3. Luehr, N., Ufimtsev, I. S. & Martinez, T. J. Dynamic precision for electron repulsion integral evaluation on graphical processing units (GPUs). *J. Chem. Theory Comput.* **7,** 949–954 (2011).
4. Kulik, H. J., Luehr, N., Ufimtsev, I. S. & Martinez, T. J. *Ab initio* quantum chemistry for protein structures. *J. Phys. Chem. B* **116,** 12501–12509 (2012).
5. Yin, Y. *et al.* Formation of hollow nanocrystals through the nanoscale Kirkendall effect. *Science* **304,** 711–714 (2004).
6. Ensing, B., De Vivo, M., Liu, Z. W., Moore, P. & Klein, M. L. Metadynamics as a tool for exploring free energy landscapes of chemical reactions. *Acc. Chem. Res.* **39,** 73–81 (2006).
7. Pietrucci, F. & Andreoni, W. Graph theory meets *ab initio* molecular dynamics: atomic structures and transformations at the nanoscale. *Phys. Rev. Lett.* **107,** 085504 (2011).
8. Iannuzzi, M., Laio, A. & Parrinello, M. Efficient exploration of reactive potential energy surfaces using Car–Parrinello molecular dynamics. *Phys. Rev. Lett.* **90,** 238302 (2003).
9. Zimmerman, P. M. Automated discovery of chemically reasonable elementary reaction steps. *J. Comput. Chem.* **34,** 1385–1392 (2013).
10. Rappoport, D., Galvin, C. J., Zubarev, D. Y. & Aspuru-Guzik, A. Complex chemical reaction networks from heuristics-aided quantum chemistry. *J. Chem. Theory Comput.* **10,** 897–907 (2014).
11. Virshup, A. M., Contreras-García, J., Wipf, P., Yang, W. & Beratan, D. N. Stochastic voyages into uncharted chemical space produce a representative library of all possible drug-like compounds. *J. Am. Chem. Soc.* **135,** 7296–7303 (2013).
12. Maeda, S. & Morokuma, K. Toward predicting full catalytic cycle using automatic reaction path search method: a case study on $HCo(CO)_3$-catalyzed hydroformylation. *J. Chem. Theory Comput.* **8,** 380–385 (2012).
13. Wales, D. J., Miller, M. A. & Walsh, T. R. Archetypal energy landscapes. *Nature* **394,** 758–760 (1998).
14. Goldman, N., Reed, E. J., Fried, L. E., Kuo, I. F. W. & Maiti, A. Synthesis of glycine-containing complexes in impacts of comets on early Earth. *Nature Chem.* **2,** 949–954 (2010).
15. Goldman, N. *et al. Ab initio* simulation of the equation of state and kinetics of shocked water. *J. Chem. Phys.* **130,** 124517 (2009).
16. Bernasconi, M., Chiarotti, G. L., Focher, P., Parrinello, M. & Tosatti, E. Solid-state polymerization of acetylene under pressure: *ab initio* simulation. *Phys. Rev. Lett.* **78,** 2008–2011 (1997).
17. Feller, D. & Peterson, K. A. An examination of intrinsic errors in electronic structure methods using the Environmental Molecular Sciences Laboratory computational results database and the Gaussian-2 set. *J. Chem. Phys.* **108,** 154–176 (1998).
18. Sousa, S. F., Fernandes, P. A. & Ramos, M. J. General performance of density functionals. *J. Phys. Chem. A* **111,** 10439–10452 (2007).
19. Harding, M. E. *et al.* High-accuracy extrapolated *ab initio* thermochemistry. III. Additional improvements and overview. *J. Chem. Phys.* **128,** 114111 (2008).
20. Miller, S. L. & Urey, H. C. Organic compound synthesis on the primitive Earth. *Science* **130,** 245–251 (1959).
21. Trout, C. C. & Badding, J. V. Solid state polymerization of acetylene at high pressure and low temperature. *J. Phys. Chem. A* **104,** 8142–8145 (2000).
22. Sakashita, M., Yamawaki H. & Aoki, K. FT-IR study of the solid state polymerization of acetylene under pressure. *J. Phys. Chem.* **100,** 9943–9947 (1996).
23. Virshup, A. M. *et al.* Photodynamics in complex environments: *ab initio* multiple spawning quantum mechanical molecular mechanical dynamics. *J. Phys. Chem. B* **113,** 3280–3291 (2009).
24. Danger, G., Plasson, R. & Pascal R. Pathways for the formation and evolution of peptides in prebiotic environments. *Chem. Soc. Rev.* **41,** 5416–5429 (2012).
25. Menten, K. M. & Wyrowski, F. in *Interstellar Molecules: Their Laboratory and Interstellar Habitat* (eds Yamada, K. M. T. & Winnewisser, G.) 27–42 (Springer Tracts in Modern Physics 241, Springer, 2011).
26. Szori, M. *et al.* Chemical evolution of biomolecule building blocks. Can thermodynamics explain the accumulation of glycine in the prebiotic ocean? *Phys. Chem. Chem. Phys.* **13,** 7449–7458 (2011).
27. Wahner, A., Mentel, T. F. & Sohn, M. Gas-phase reaction of N_2O_5 with water vapor: importance of heterogeneous hydrolysis of N_2O_5 and surface desorption of HNO_3 in a large Teflon chamber. *Geophys. Res. Lett.* **25,** 2169–2172 (1998).
28. Kasting, J. F. Earth's early atmosphere. *Science* 1993, **259,** 920–926.
29. Cleaves, H. J., Chalmers, J. H., Lazcano, A., Miller, S. L. & Bada J. L. A reassessment of prebiotic organic synthesis in neutral planetary atmospheres. *Origins Life Evol. Biosph.* **38,** 105–115 (2008).
30. Isborn, C. M., Luehr, N., Ufimtsev, I. S. & Martinez, T. J. Excited-state electronic structure with configuration interaction singles and Tamm–Dancoff time-dependent density functional theory on graphical processing units. *J. Chem. Theory Comput.* **7,** 1814–1823 (2011).
31. Titov, A. V., Ufimtsev, I. S., Luehr, N. & Martinez, T. J. Generating efficient quantum chemistry codes for novel architectures. *J. Chem. Theory Comput.* **9,** 213–221 (2013).
32. Ufimtsev, I. S. & Martinez, T. J. Graphical processing units for quantum chemistry. *Comput. Sci. Eng.* **10,** 26–34 (2008).
33. Ufimtsev, I. S. & Martinez, T. J. Quantum chemistry on graphical processing units. 1. Strategies for two-electron integral evaluation. *J. Chem. Theory Comput.* **4,** 222–231 (2008).
34. Ufimtsev, I. S. & Martinez, T. J. Quantum chemistry on graphical processing units. 2. Direct self-consistent-field implementation. *J. Chem. Theory Comput.* **5,** 1004–1015 (2009).
35. Saunders, V. R. & Hillier, I. H. Level-shifting method for converging closed-shell Hartree–Fock wavefunctions. *Int. J. Quantum Chem.* **7,** 699–705 (1973).
36. Hu, X. & Yang, W. Accelerating self-consistent field convergence with the augmented Roothaan–Hall energy function. *J. Chem. Phys.* **132,** 054109 (2010).
37. Pulay, P. Convergence acceleration of iterative sequences—the case of SCF iteration. *Chem. Phys. Lett.* **73,** 393–398 (1980).
38. Hagberg, A. A., Schult, D. A. & Swart, P. J. in *Proceedings of the 7th Python in Science Conference* (eds Varoquaux, G., Vaught. T & Millman, J.) 11–15 (SciPy, 2008).
39. Pedregosa F. *et al.* Scikit-learn: machine learning in Python. *J. Mach. Learn. Res.* **12,** 2825–2830 (2011).
40. Becke, A. D. Density-functional thermochemistry. 3. The role of exact exchange. *J. Chem. Phys.* **98,** 5648–5652 (1993).
41. Guner, V. *et al.* A standard set of pericyclic reactions of hydrocarbons for the benchmarking of computational methods: the performance of *ab initio*, density functional, CASSCF, CASPT2, and CBS-QB3 methods for the prediction of activation barriers, reaction energetics, and transition state geometries. *J. Phys. Chem. A* **107,** 11445–11459 (2003).
42. Swart, M., Sola, M. & Bickelhaupt, F. M. Energy landscapes of nucleophilic substitution reactions: a comparison of density functional theory and coupled cluster methods. *J. Comput. Chem.* **28,** 1551–1560 (2007).
43. Van Voorhis, T. and Head-Gordon, M. Benchmark variational coupled cluster doubles results. *J. Chem. Phys.* **113,** 8873–8879 (2000).
44. Zhang, J. and Valeev, E. F. Prediction of reaction barriers and thermochemical properties with explicitly correlated coupled-cluster methods: a basis set assessment. *J. Chem. Theory Comput.* **8,** 3175–3186 (2012).
45. Peters, B., Heyden, A., Bell, A. T. & Chakraborty, A. A growing string method for determining transition states: comparison to the nudged elastic band and string methods. *J. Chem. Phys.* **120,** 7877–7886 (2004).
46. Shao, Y. *et al.* Advances in methods and algorithms in a modern quantum chemistry program package. *Phys. Chem. Chem. Phys.* **8,** 3172–3191 (2006).
47. Bui, P., Rajan, D., Abdul-Wahid, B., Izaguirre, J. & Thain, D. Work Queue + Python: a framework for scalable scientific ensemble applications. Workshop on Python for High Performance and Scientific Computing (PyHPC, 2011).

Acknowledgements

This work was supported by the National Science Foundation (OCI-1047577), the National Institutes of Health (U54 GM072970) and the Department of Defense through a National Security Science and Engineering Faculty Fellowship from the Office of the Assistant Secretary of Defense for Research and Engineering. This work included calculations performed on the Blue Waters supercomputer at the National Centre for Supercomputing Applications and funded by the National Science Foundation's Office of Cyber Infrastructure. Further computational support was provided by the AMOS program within the Chemical Sciences, Geosciences and Biosciences Division of the Office of Basic Energy Sciences, Office of Science, Department of Energy. We are grateful to E. G. Hohenstein, N. Luehr, S. D. Fried, S. Izmailov, Y. Zhao and C-Y. Wang for helpful suggestions.

Author contributions

L-P.W., A.T., F.L. and T.J.M. designed the nanoreactor simulation studies. L-P.W., R.M., V.S.P. and T.J.M. designed the energy refinement and network analysis. L.P.W. carried out the simulations and analysis. L-P.W., V.S.P. and T.J.M. co-wrote the manuscript. All authors discussed the results and commented on the manuscript.

Additional information

Supplementary information is available in the online version of the paper. Reprints and permissions information is available online at www.nature.com/reprints. Correspondence and requests for materials should be addressed to T.J.M.

Competing financial interests

The authors declare no competing financial interests.