JCIM JOURNAL OF CHEMICAL INFORMATION AND MODELING

pubs.acs.org/jcim

Viewpoint

Do Large Language Models Understand Chemistry? A Conversation with ChatGPT

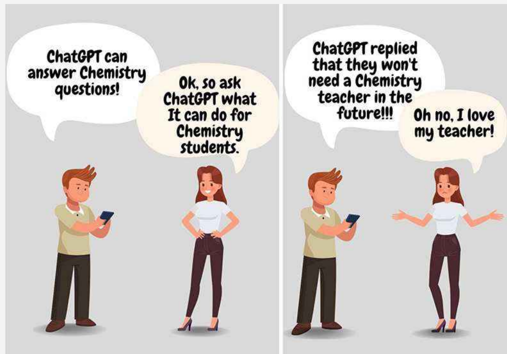
Cayque Monteiro Castro Nascimento and André Silva Pimentel*

Cite This: *J. Chem. Inf. Model.* 2023, 63, 1649−1655

Read Online

ACCESS | Metrics & More | Article Recommendations

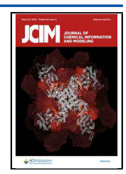
ABSTRACT: Large language models (LLMs) have promised a revolution in answering complex questions using the ChatGPT model. Its application in chemistry is still in its infancy. This viewpoint addresses the question of how well ChatGPT understands chemistry by posing five simple tasks in different subareas of chemistry.



Large language models (LLMs) are tools of artificial intelligence (AI) that use machine learning algorithms to generate text. This allows them to guess or predict words and create phrases, paragraphs, and a full essay that reflect how humans write and speak. These tasks are performed with unbelievable proficiency in different languages.[1−3] LLMs have recently remade natural language processing because they have helped to develop robust trained models for many different tasks using data sets with even billions of words.[4−7] These models are also trained using complex algorithms to recognize word-based patterns, allowing them to learn its context and natural language.[8,9] LLMs are input with text excerpts that are partially masked or obscured. Thus, a neural network attempts to predict the absent elements and then evaluates the prediction to the original text. The neural network executes this task iteratively adjusting parameters based on the output. Finally, this neural network builds a model of how words are related to each other in sentences.[10−13] LLMs are powerful and versatile tools for many applications that might include understanding chemistry, perhaps.[14−20]

Examples of LLMs are as follows: (1) Turing NLGm was released and developed by Microsoft in early 2020 with the largest amount of data to date, 17-billion parameters.[21] (2) Gopher is a model with 280-billion-parameters developed by DeepMind.[22] It excels in STEM disciplines, which is the first clue that LLM can understand chemistry. (3) The GPT-3 model was developed by the AI research and deployment company OpenAI. It is nurtured with a publicly available data set of around 570 gigabytes of text information. GPT-3 has released one of the largest neural networks that can reconstruct merely anything with a language structure, including a computer code.[20,23] (4) Galactica is a huge open-source language model designed to help scientists and was released by Meta, hoping to clean up its image from criticism.[24] After just three days, Meta decided to not demonstrate its model to the public, supposedly because it might not work as well as users want or not work with the necessary ethics, receiving more criticism.[25] Critics comment that there is a blind spot to the serious limitations of the big language models, especially because the data used to build the models are sometimes not curated, freely available anywhere on the web, or even from secondary literature.[25] From this point of view, if it is asked about a physical or chemical property, a LLM can answer any value or property about a chemical compound, wrong or correct, because the data are not curated or from secondary literature. Therefore, it is important to remember that LLMs have limited understanding of the text it analyzes or generates. If LLMs capture a wrong value in its training, it might answer this value after being asked. The generated answers may be apparently valid, but LLMs do not have the ability to reason or demonstrate understanding about the subject. LLMs are also not able to respond about future trends in chemistry. There are also several other limitations.[24,25]

Received: February 23, 2023
Published: March 16, 2023



https://doi.org/10.1021/acs.jcim.3c00285
J. Chem. Inf. Model. 2023, 63, 1649−1655

The aim of this viewpoint is to raise simple tasks that are answered correctly or incorrectly, precise or not, using LLMs in chemistry. Criticisms remind us that the only thing LLMs "know" for sure is how words and sentences are formed. Everything else is speculation. Is this criticism too strong? It is important to mention that what is speculation now might not be speculation in the future when LLMs will be better built using more reliable data.[26] In this viewpoint we briefly review some of the shortcomings still present in LLMs available today with the intent to bring the attention of users and developers to the need for advancements.

To illustrate the underlying issues, we focus our discussion on specific tasks that LLMs might apply in chemistry using the OpenAI ChatGPT with the InstructGPT model, text-davinci-003, which has knowledge of chemistry equations and common calculations.[27] However, the outcomes might not be of relevance to other LLMs described anywhere. It follows the control parameters used in the predictions made in this viewpoint. Temperature is one of the most important settings to control the output of the GPT-3 engine. It controls the randomness of the generated text.[28] A value of 0 makes the engine deterministic, which means that it will always generate the same output for a given input text, using 0.1 will be more deterministic. The maximum tokens are 256 (1 token is around four characters) that can be generated by the model.[29] A standard "top p" parameter equal to 1 controls how many different words or phrases the language model considers when it is trying to predict a sentence. A frequency penalty of 0 was used to lower the chances of a word being selected again. Also used was a presence penalty of 0 that encourages the model to make novel predictions.

FIRST TASK: CONVERT A COMPOUND NAME INTO THE SMILES CHEMICAL REPRESENTATION AND VICE VERSA

The attempt to convert a compound name into the SMILES chemical representation used the following question: What is the SMILES representation of {compound name}?[30,31]

The attempt to convert SMILES into a compound name used the following question: What is the compound name whose the SMILES representation is {SMILES}?

The conversion of compound names to SMILES chemical representations,[30] and vice versa, is a difficult task for LLMs even for the case of simple alkanes and alkenes (Table 1). The hit rate is around 27% for both tasks. It is challenging even for very small molecules such as alkanes of two or three carbon atoms. It is also difficult for IUPAC or common names of some compounds. For larger straight chain, branched, cyclic, or aromatic hydrocarbon compounds of 4 to 10 carbon atoms, the ChatGPT model makes a lot of confusion as observed in Table 1. It does not comprehend the difference between alkanes and alkenes, benzene and cyclohexene, two alkanes with the difference of several carbon atoms, *cis* and *trans* isomers, and so on. It is also interesting that it adds halogen or oxygen atoms that do not exist in the molecule. More robust string representations may be more suitable, but a couple attempts showed that the ChatGPT model does not understand the actual and robust SELFIES representation, for example.[32,33]

Table 1. Conversion of Compound Names into the SMILES Chemical Representations, and Vice Versa[a]

Convert compound name into the SMILES chemical representation (6/23=26%)			
Name	SMILES	Name	SMILES
Ethane	C=C	Ethene	C=C
Propane	C(C)(C)C	Propene Propylene	C=C C=C(C)C
Butane	CCCCC	But-2-ene	C=C(C)C
Pentane	CCCCC	cis-but-2-ene trans-but-2-ene	C/C=C/C C/C=C/C
Hexane	CCCCC	2-Methyl-2-butene	C=C(C)C(C)C
Heptane	CCCCCCC	2-Methyl-1-butene	C=C(C)C(C)(C)C
Octane	CCCCCCC	Isoprene 2-methyl-1,3-butadiene	C=C(C)C(C)C C=C(C)C(C)=C
Nonane	CCCCCCCCCC	Cyclohexane	C1CCCCC1
Decane	CCCCCCCCCCC	Cyclohexene	C1=CC=CC=C1
Xylene	C1=CC=CC=C1	Benzene	C1=CC=CC=C1
Convert SMILES into the compound name (5/18=28%)			
SMILES	Name	SMILES	Name
CC	Ethane	C=C	Ethene
CCC	Tricarbon Monoxide	C=CC CC=C	Ethylene Ethene
CCCC	Butane	CC=CC	1,3-Butadiene
CCCCC	Pentane	C/C=C\C C/C=C/C	1,3-Dichloropropene 1,3-Butadiene
CCCCCC	Hexane	CC=C(C)C	2-Buten-1-ol
CCCCCC CC	Hexane	CCC(=C)C	3-Methylbut-1-ene
CCCCCC CCC	Octane	CC(=C)C=C	1,3-Butadiene
CCCCCC CCCC	Octane	C1=CC=CC=C1	Cyclopentene

[a]Green represent correct answers.[30,31]

SECOND TASK: FINDING INFORMATION ON OCTANOL–WATER PARTITION COEFFICIENTS OF CHEMICAL COMPOUNDS

The attempt to find octanol–water partition coefficients of essential oil components was made using the following question: What is the octanol–water partition coefficient of the {compound name}?[34,35]

The experimental octanol–water partition coefficients ($\log P_{\mathrm{exp}}$) of essential oil components is compared with what ChatGPT finds in the literature ($\log P_{\mathrm{ChatGPT}}$) (Table 2). It is important to mention that the experimental techniques used to measure this property for hydrophobic molecules are not the standard ones.[36−38] The ChatGPT model found reasonable values for this property, sometimes much better than those found using biocheminformatics tools. Excluding the unknown octanol–water partition coefficients (ChatGPT model answers the octanol–water partition coefficient as unknown for some compounds), the mean relative error was around 31% which is very reasonable for these kinds of complex molecules.

THIRD TASK: GETTING STRUCTURAL INFORMATION ON COORDINATION COMPOUNDS

The attempt to find the geometries of coordination compounds was performed using the following question: What is the geometry of the coordination compound {compound}?[39]

Table 2. Comparison between Experimental Octanol−Water Partition Coefficients ($\log P_{exp}$) of Essential Oil Components and Value of the Same Property Found Using the ChatGPT Model ($\log P_{ChatGPT}$)[a]

Name	$\log P_{exp}$	$\log P_{ChatGPT}$	Error (%)
linalool	3.19	4.3	35
limonene	4.38	4.7	7
gamma-Terpinene	4.36	4.2	-4
(+)-Camphene	4.22	UNK	
camphor	2.74	4.3	57
terpinyl acetate	3.96	UNK	-
eugenol	2.23	4.6	106
citronellal	3.83	4.2	10
citronellol	3.91	4.3	10
p-cymene	4.1	4.2	2
(R)-(-)-Carvone	2.47	3.3	34
(1R)-(-)-fenchone	2.59	4.7	81
geraniol	2.54	4.5	77
carvacrol	2.99	4.4	47
thymol	3.15	4.3	37
alpha-terpinene	4.25	4.2	-1
(1R)-Camphor	2.41	4.2	74
(±)-β-citronellol	3.21	4.3	34
Eucalyptol	2.89	4.2	45
L-(-)-menthol	3.85	4.2	9
(-)-menthone	3.33	4.3	29
p-Ment-6-en-2,8-diol	1.81	UNK	-
(-)-cis-Myrtanylamine	2.05	UNK	-
(1S)-(-)-Verbenone	2.23	3.2	43
Linalool oxide	2.43	UNK	-
Piperitone	2.85	UNK	-
1,4-Cineole	2.97	4.7	58
Myrtenal	2.98	UNK	-
(-)-Borneol	3.01	4.3	43
Dihydrocarvone	3.08	UNK	-
(+)-Pulegone	3.08	4.3	40
(-)-Carveol	3.12	4.2	35
(-)-Perilla aldehyde	3.13	UNK	-
(S)-cis-Verbenol	3.16	4.3	36
(S)-(-)-Perilla alcohol	3.17	UNK	-
(+)-Fenchol	3.17	4.2	32
(+)-Isomenthol	3.19	2.6	-18
Limonene oxide	3.2	UNK	-
Dihydrocarveol	3.21	UNK	-
(+/-)-Isoborneol	3.24	4.3	33
Terpinen-4-ol	3.26	UNK	-
alpha-Terpineol	3.28	4.3	31
Methyleugenol	3.45	4.2	22
Nerol	3.47	UNK	-
beta-Ionone	3.84	4.3	12
alpha-Ionone	3.85	4.2	9
Linalyl acetate	3.93	4.3	9
Neryl acetate	3.98	UNK	-
Menthyl acetate	4	UNK	-
Geranyl acetate	4.04	4.2	4
alpha-Terpinolene	4.24	UNK	-
Car-3-ene	4.38	UNK	-
Car-2-ene	4.44	UNK	-
(+)-alpha-Pinene	4.44	4.3	-3

[a]The relative error (%) is also presented for each compound. Values in pink are obtained from ref 34 and blue from ref 35. UNK: unknown.

The geometries of coordination compounds with coordination numbers from 2 to 12 are predicted in Table 3. The ChatGPT model makes the right prediction in 5 of 12 coordination compounds. It is important to mention that it predicts almost correctly two coordination compounds, $K_3[NbOF_6]$ and $(NH_4)_2Ce(NO_3)_6$. Both compounds are only a different kind of octahedron. So, if this is considered correct, the hit rate of the ChatGPT model is 58%, which is considered good because some of these compounds are not common.

FOURTH TASK: WATER SOLUBILITY OF POLYMERS

The attempt to find the water solubility of polymers was made using the following question: What is the water solubility of {polymer}?[40]

The ChatGPT model makes the correct prediction of the water solubilities of 11 polymers (Table 4) because they have important applications in industry and academy. It is honest to note that this task is simple to predict even for a student because the chemical structure and functional group of the monomer are clear and simple evidence for a student to make a good prediction. Also, it is important to mention that the ChatGPT prediction is more reliable when the question is about something contextualized.

FIFTH TASK: MOLECULAR POINT GROUPS

The attempt to find the molecular point groups of the molecules was performed using the following question: What is the molecular point group of {molecule}?[41]

The ChatGPT model makes the right prediction in 6 of 10 molecular point groups of simple molecules compounds (Table 5). This is considered reasonable because this subject is not as popular and common as coordination chemistry, for example. If complex molecules are used in these questions as used for the coordination chemistry subject, the hit rate might be less than 60%.

CONCLUDING REMARKS

LLMs are nowadays applied to interpret questions in chemistry subjects and answer them to understand if LLMs can comprehend chemistry. Although researchers have recently stated that they found high accuracy on chemistry questions using some tricks,[31] it is presented here in five tasks that the accuracy in answering the questions was between 25% and 100% without any tricks. The low or high accuracy depends on several important considerations: reasonable prompts should give correct answers, questions on popular subjects are easily answered, very specific topics that are not well included in a database or are not well trained in the model gives low accuracy, and the development of better prompts or strategies for training and fitting this knowledge in models might output better results.[13]

In this viewpoint, we attempted to mimic a regular student prompting the ChatGPT model to answer questions on chemistry subjects without using any tricks such as inserting copyright notices in source files or fine-tuning with human feedback. Although aligning language models with human intent is a promising direction to get correct answers,[13,31] nevertheless, it is important that care must be taken using completions with difficult prompts.[26] It is also noted that LLMs always answer something. It is somewhat interesting and

https://doi.org/10.1021/acs.jcim.3c00285
J. Chem. Inf. Model. 2023, 63, 1649−1655

Table 3. Geometries of Coordination Compounds with Coordination Numbers from 2 to 12[a]

Coordination compound	Geometry	Correct?
$[Ag(NH_3)_2]^+$	Linear	Yes
$[Cu(CN)_3]^{2-}$	Trigonal planar	Yes
$CoCl_2(pyr)_2$	Square planar	No. Tetrahedral
cis-$PtCl_2(NH_3)_2$	Square planar	Yes
$VO(acac)_2$	Octahedral	No. Square pyramid
$[Ni(CN)_5]^{3+}$ $[Cr(en)_3][Ni(CN)_5]$	Trigonal bipyramidal Octahedral	Yes No. Trigonal bipyramidal
$[Co(en)_3]Cl_3$	Octahedral	Yes
$K_3[NbOF_6]$	Octahedral	Almost. Capped octahedral
bis-(tert-butylacac)$_2$(DMSO)di-oxoUranium	Octahedral	No. Pentagonal Bipyramid
$(NH_4)_2Ce(NO_3)_6$ Ceric ammonium nitrate	Octahedral Tetrahedral	Almost. Cuboctahedron No. Cuboctahedron

[a]Green represents correct answers.[39] Ligand abbreviations: pyr = pyruvic acid, acac = acetylacetonate, en = ethylenediamine, and DMSO = dimethyl sulfoxide.

Table 4. Water Solubility of 11 Polymers[a]

Polymer	Answer
Poly(ethylene glycol)	Poly(ethylene glycol) with a molecular weight of less than 1,000 is highly soluble in water, while those with a molecular weight of more than 10,000 are only slightly soluble.
Polyvinyl pyrrolidone	The water solubility of polyvinyl pyrrolidone (PVP) is generally considered to be very high, with a solubility of up to 20% in water at room temperature.
Polyvinyl alcohol	The water solubility of polyvinyl alcohol (PVA) is generally considered to be very high, with most grades of PVA being completely soluble in cold water.
Polyacrylic acid	The water solubility of polyacrylic acid is very high, with a solubility of up to 90% in water at room temperature.
Polyacrylamide	Generally, anionic polyacrylamide is more soluble in water than cationic polyacrylamide. Anionic polyacrylamide is typically soluble in water up to 1-2%, while cationic polyacrylamide is usually only soluble up to 0.1-0.2%.
N-(2-Hydroxypropyl) methacrylamide	The water solubility of N-(2-Hydroxypropyl) methacrylamide is approximately 0.1 g/L at 25 °C.
Polyethylene	Polyethylene is insoluble in water.
Polyvinylchloride	The water solubility of polyvinyl chloride (PVC) is very low, with only 0.01-0.02 g/100 mL of water at room temperature.
Polyvinyl acetate	The water solubility of polyvinyl acetate is very low, with only 0.1 g/L at 20 °C.
Polystyrene	Polystyrene is not water soluble.
Polydimethylsiloxane	Polydimethylsiloxane is insoluble in water.

[a]100% of answers are correctly predicted.[40]

surprising that non-English languages may generate even better outputs.

It is disappointing that the conversion of SMILES representation into compound names, and vice versa, gives such low accuracy. This prompt gives a low hit rate because the answer sometimes misses or adds a methyl group, for example. Other confusions are including a nonexistent atom in the representation, confounding between regular cyclic and aromatic, or not understanding isomers. Unfortunately, the ChatGPT model did not perform well in several attempts to understand the most actual and robust string representation, SELFIES.[32,33]

Finally, it is essential to give a rationalist prediction of the future in the field of LLMs. It is important to mention that neither experimental nor computational chemists should fear the development of LLMs. The automation of tasks should not reduce the need for hands and creativity of experimental or computational chemists. Instead, many exciting and better artificial intelligence tools[42−44] should be integrated into research to expand and solve even complex problems that challenge researchers for long periods of time, reducing the effort and facilitating the resolution of the problem.

https://doi.org/10.1021/acs.jcim.3c00285
J. Chem. Inf. Model. 2023, 63, 1649−1655

Table 5. Molecular Point Groups of Simple Molecules[a]

Molecule	Point group	Correct?
H_2O	C_{2v}	Yes
CH_4	C_{3v}	No. It is T_d
NH_3	C_{3v}	Yes
Ethene	C_{2v}	No. It is D_{2h}
Acetylene	$D_{\infty h}$	Yes
Benzene	D_{6h}	Yes
AsF_5	C_{3v}	No. It is D_{3h}
SCl_4	C_{4v}	No. It is C_{2v}.
BrF_3	C_{2v}	Yes
Ferrocene	D_{5h}	Yes

[a]Green represents are correct answers[41].

ASSOCIATED CONTENT

Data Availability Statement

The code is executed in Python 3.8.10 and available at https://github.com/andresilvapimentel/AI4Chem. Access to OpenAI GPT-3 InstructGPT (text-davinci-003) is governed by OpenAI and not the authors.

AUTHOR INFORMATION

Corresponding Author

André Silva Pimentel − *Departamento de Química, Pontifícia Universidade Católica do Rio de Janeiro, Rio de Janeiro, RJ 22451-900, Brazil;* orcid.org/0000-0002-1301-0561; Email: a_pimentel@puc-rio.br

Author

Cayque Monteiro Castro Nascimento − *Departamento de Química, Pontifícia Universidade Católica do Rio de Janeiro, Rio de Janeiro, RJ 22451-900, Brazil;* orcid.org/0000-0003-4712-0598

Complete contact information is available at: https://pubs.acs.org/10.1021/acs.jcim.3c00285

Notes

The authors declare no competing financial interest.

ACKNOWLEDGMENTS

The authors are thankful to the research productivity fellowship granted by CNPq (310166/2020-9), the INCT-FCx (CNPq Grants 573560/2008-0, 465259/2014-6, and 302554/2017-3, and FAPESP Grant 2014/50983-3), and CAPES (Finance Code 001). The authors also acknowledge the FAPERJ NanoHealth Research Network (E-26/010.000983/2019), the FAPERJ Program for Thematic Projects in the State of Rio de Janeiro (210.104/2020), and the FAPERJ award "Scientist of Our State" (201.186/2022).

REFERENCES

(1) Bird, S.; Klein, E.; Loper, E. *Natural Language Processing with Python: Analyzing Text with the Natural Language Toolkit*, 1st ed.; O'Reilly Media, 2009; Vol. *1*.

(2) Tunstall, L.; von Werra, L.; Wolf, T. *Natural Language Processing with Transformers: Building Language Applications with Hugging Face*, 1st ed.; O'Reilly Media, 2022; Vol. *1*.

(3) Kublik, S.; Saboo, S. *GPT-3: Building Innovative NLP Products Using Large Language Models*, 1st ed.; O'Reilly Media, 2022; Vol. *1*.

(4) Hocky, G. M.; White, A. D. Natural Language Processing Models That Automate Programming Will Transform Chemistry Research and Teaching. *Digit Discovery* **2022**, *1*, 79−83.

(5) Srivastava, A.; Rastogi, A.; Rao, A.; Shoeb, A. A. M.; Abid, A.; Fisch, A.; Brown, A. R.; Santoro, A.; Gupta, A.; Garriga-Alonso, A.; Kluska, A.; Lewkowycz, A.; Agarwal, A.; Power, A.; Ray, A.; Warstadt, A.; Kocurek, A. W.; Safaya, A.; Tazarv, A.; Xiang, A.; Parrish, A.; Nie, A.; Hussain, A.; Askell, A.; Dsouza, A.; Slone, A.; Rahane, A.; Iyer, A. S.; Andreassen, A.; Madotto, A.; Santilli, A.; Stuhlmüller, A.; Dai, A.; La, A.; Lampinen, A.; Zou, A.; Jiang, A.; Chen, A.; Vuong, A.; Gupta, A.; Gottardi, A.; Norelli, A.; Venkatesh, A.; Gholamidavoodi, A.; Tabassum, A.; Menezes, A.; Kirubarajan, A.; Mullokandov, A.; Sabharwal, A.; Herrick, A.; Efrat, A.; Erdem, A.; Karakaş, A.; Roberts, B. R.; Loe, B. S.; Zoph, B.; Bojanowski, B.; Özyurt, B.; Hedayatnia, B.; Neyshabur, B.; Inden, B.; Stein, B.; Ekmekci, B.; Lin, B. Y.; Howald, B.; Diao, C.; Dour, C.; Stinson, C.; Argueta, C.; Ramírez, C. F.; Singh, C.; Rathkopf, C.; Meng, C.; Baral, C.; Wu, C.; Callison-Burch, C.; Waites, C.; Voigt, C.; Manning, C. D.; Potts, C.; Ramirez, C.; Rivera, C. E.; Siro, C.; Raffel, C.; Ashcraft, C.; Garbacea, C.; Sileo, D.; Garrette, D.; Hendrycks, D.; Kilman, D.; Roth, D.; Freeman, D.; Khashabi, D.; Levy, D.; González, D. M.; Perszyk, D.; Hernandez, D.; Chen, D.; Ippolito, D.; Gilboa, D.; Dohan, D.; Drakard, D.; Jurgens, D.; Datta, D.; Ganguli, D.; Emelin, D.; Kleyko, D.; Yuret, D.; Chen, D.; Tam, D.; Hupkes, D.; Misra, D.; Buzan, D.; Mollo, D. C.; Yang, D.; Lee, D.-H.; Shutova, E.; Cubuk, E. D.; Segal, E.; Hagerman, E.; Barnes, E.; Donoway, E.; Pavlick, E.; Rodola, E.; Lam, E.; Chu, E.; Tang, E.; Erdem, E.; Chang, E.; Chi, E. A.; Dyer, E.; Jerzak, E.; Kim, E.; Manyasi, E. E.; Zheltonozhskii, E.; Xia, F.; Siar, F.; Martínez-Plumed, F.; Happé, F.; Chollet, F.; Rong, F.; Mishra, G.; Winata, G. I.; de Melo, G.; Kruszewski, G.; Parascandolo, G.; Mariani, G.; Wang, G.; Jaimovitch-López, G.; Betz, G.; Gur-Ari, G.; Galijasevic, H.; Kim, H.; Rashkin, H.; Hajishirzi, H.; Mehta, H.; Bogar, H.; Shevlin, H.; Schütze, H.; Yakura, H.; Zhang, H.; Wong, H. M.; Ng, I.; Noble, I.; Jumelet, J.; Geissinger, J.; Kernion, J.; Hilton, J.; Lee, J.; Fisac, J. F.; Simon, J. B.; Koppel, J.; Zheng, J.; Zou, J.; Kocoń, J.; Thompson, J.; Kaplan, J.; Radom, J.; Sohl-Dickstein, J.; Phang, J.; Wei, J.; Yosinski, J.; Novikova, J.; Bosscher, J.; Marsh, J.; Kim, J.; Taal, J.; Engel, J.; Alabi, J.; Xu, J.; Song, J.; Tang, J.; Waweru, J.; Burden, J.; Miller, J.; Balis, J. U.; Berant, J.; Frohberg, J.; Rozen, J.; Hernandez-Orallo, J.; Boudeman, J.; Jones, J.; Tenenbaum, J. B.; Rule, J. S.; Chua, J.; Kanclerz, K.; Livescu, K.; Krauth, K.; Gopalakrishnan, K.; Ignatyeva, K.; Markert, K.; Dhole, K. D.; Gimpel, K.; Omondi, K.; Mathewson, K.; Chiafullo, K.; Shkaruta, K.; Shridhar, K.; McDonell, K.; Richardson, K.; Reynolds, L.; Gao, L.; Zhang, L.; Dugan, L.; Qin, L.; Contreras-Ochando, L.; Morency, L.-P.; Moschella, L.; Lam, L.; Noble, L.; Schmidt, L.; He, L.; Colón, L. O.; Metz, L.; Şenel, L. K.; Bosma, M.; Sap, M.; ter Hoeve, M.; Farooqi, M.; Faruqui, M.; Mazeika, M.; Baturan, M.; Marelli, M.; Maru, M.; Quintana, M. J. R.; Tolkiehn, M.; Giulianelli, M.; Lewis, M.; Potthast, M.; Leavitt, M. L.; Hagen, M.; Schubert, M.; Baitemirova, M. O.; Arnaud, M.; McElrath, M.; Yee, M. A.; Cohen, M.; Gu, M.; Ivanitskiy, M.; Starritt, M.; Strube, M.; Swędrowski, M.; Bevilacqua, M.; Yasunaga, M.; Kale, M.; Cain, M.; Xu, M.; Suzgun, M.; Tiwari, M.; Bansal, M.; Aminnaseri, M.; Geva, M.; Gheini, M.; T, M. V.; Peng, N.; Chi, N.; Lee, N.; Krakover, N. G.-A.; Cameron, N.; Roberts, N.; Doiron, N.; Nangia, N.; Deckers, N.; Muennighoff, N.; Keskar, N. S.; Iyer, N. S.; Constant, N.; Fiedel, N.; Wen, N.; Zhang, O.; Agha, O.; Elbaghdadi, O.; Levy, O.; Evans, O.; Casares, P. A. M.; Doshi, P.; Fung, P.; Liang, P. P.; Vicol, P.; Alipoormolabashi, P.; Liao, P.; Liang, P.; Chang, P.; Eckersley, P.; Htut, P. M.; Hwang, P.; Miłkowski, P.; Patil, P.; Pezeshkpour, P.; Oli, P.; Mei, Q.; Lyu, Q.; Chen, Q.; Banjade, R.; Rudolph, R. E.; Gabriel, R.; Habacker, R.; Delgado, R. R.; Millière, R.; Garg, R.; Barnes, R.; Saurous, R. A.; Arakawa, R.; Raymaekers, R.; Frank, R.; Sikand, R.; Novak, R.; Sitelew, R.; LeBras, R.; Liu, R.; Jacobs, R.; Zhang, R.; Salakhutdinov, R.; Chi, R.; Lee, R.; Stovall, R.; Teehan, R.; Yang, R.; Singh, S.; Mohammad, S. M.; Anand, S.; Dillavou, S.; Shleifer, S.; Wiseman, S.; Gruetter, S.; Bowman, S. R.; Schoenholz, S. S.; Han, S.; Kwatra, S.; Rous, S. A.; Ghazarian, S.; Ghosh, S.; Casey, S.; Bischoff, S.; Gehrmann, S.; Schuster, S.; Sadeghi, S.; Hamdan, S.; Zhou, S.; Srivastava, S.; Shi, S.; Singh, S.; Asaadi, S.; Gu, S. S.; Pachchigar, S.; Toshniwal, S.; Upadhyay, S.; Shyamolima; Debnath; Shakeri, S.; Thormeyer, S.; Melzi, S.; Reddy, S.; Makini, S. P.; Lee, S.-H.; Torene, S.; Hatwar, S.; Dehaene, S.; Divic, S.; Ermon, S.; Biderman, S.; Lin, S.; Prasad, S.; Piantadosi, S. T.; Shieber, S. M.;

https://doi.org/10.1021/acs.jcim.3c00285
J. Chem. Inf. Model. 2023, 63, 1649−1655

Misherghi, S.; Kiritchenko, S.; Mishra, S.; Linzen, T.; Schuster, T.; Li, T.; Yu, T.; Ali, T.; Hashimoto, T.; Wu, T.-L.; Desbordes, T.; Rothschild, T.; Phan, T.; Wang, T.; Nkinyili, T.; Schick, T.; Kornev, T.; Telleen-Lawton, T.; Tunduny, T.; Gerstenberg, T.; Chang, T.; Neeraj, T.; Khot, T.; Shultz, T.; Shaham, U.; Misra, V.; Demberg, V.; Nyamai, V.; Raunak, V.; Ramasesh, V.; Prabhu, V. U.; Padmakumar, V.; Srikumar, V.; Fedus, W.; Saunders, W.; Zhang, W.; Vossen, W.; Ren, X.; Tong, X.; Zhao, X.; Wu, X.; Shen, X.; Yaghoobzadeh, Y.; Lakretz, Y.; Song, Y.; Bahri, Y.; Choi, Y.; Yang, Y.; Hao, Y.; Chen, Y.; Belinkov, Y.; Hou, Y.; Hou, Y.; Bai, Y.; Seid, Z.; Zhao, Z.; Wang, Z.; Wang, Z. J.; Wang, Z.; Wu, Z. Beyond the Imitation Game: Quantifying and Extrapolating the Capabilities of Language Models. *arXiv Preprint*, arXiv:2206.04615, 2022. https://arxiv.org/abs/2206.04615 (accessed Feb 27, 2023).

(6) Wang, S.; Guo, Y.; Wang, Y.; Sun, H.; Huang, J. SMILES-BERT. In *Proceedings of the 10th ACM International Conference on Bioinformatics, Computational Biology and Health Informatics*; Association for Computing Machinery, New York, September 7−10,2019; pp 429−436.

(7) Chithrananda, S.; Grand, G.; Ramsundar, B. ChemBERTa: Large-Scale Self-Supervised Pretraining for Molecular Property Prediction. *arXiv Preprint*, arXiv:2010.09885v2, 2020. https://arxiv.org/abs/2010.09885 (accessed Feb 27, 2023).

(8) Frey, N.; Soklaski, R.; Axelrod, S.; Samsi, S.; Gomez-Bombarelli, R.; Coley, C.; Gadepally, V. Neural Scaling of Deep Chemical Models. *ChemRxiv Preprint*, 2022. https://chemrxiv.org/engage/chemrxiv/article-details/627bddd544bdd532395fb4b5 (accessed Feb 27, 2023).

(9) Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, L.; Polosukhin, I. Attention Is All You Need.. *arXiv Preprint*, arXiv:1706.03762v5, 2017. https://arxiv.org/abs/1706.03762 (accessed Feb 27, 2023).

(10) Lane, H.; Hapke, H.; Howard, C. *Natural Language Processing in Action: Understanding, Analyzing, and Generating Text with Python*, 1st ed.; Manning Publications, 2019; Vol. *1*.

(11) Devlin, J.; Chang, M.-W.; Lee, K.; Toutanova, K. BERT: Pre-Training of Deep Bidirectional Transformers for Language Understanding. *arXiv Preprint*, arXiv:1810.04805v2, 2019. https://arxiv.org/abs/1810.04805 (accessed Feb 27, 2023).

(12) Bahdanau, D.; Cho, K.; Bengio, Y. Neural Machine Translation by Jointly Learning to Align and Translate. *arXiv Preprint*, arXiv:1409.0473v7, 2016. https://arxiv.org/abs/1409.0473 (accessed Feb 27, 2023).

(13) Ouyang, L.; Wu, J.; Jiang, X.; Almeida, D.; Wainwright, C. L.; Mishkin, P.; Zhang, C.; Agarwal, S.; Slama, K.; Ray, A.; Schulman, J.; Hilton, J.; Kelton, F.; Miller, L.; Simens, M.; Askell, A.; Welinder, P.; Christiano, P.; Leike, J.; Lowe, R. Training Language Models to Follow Instructions with Human Feedback. *arXiv Preprint*, arXiv:2203.02155v1, 2022. https://arxiv.org/abs/2203.02155 (accessed Feb 27, 2023).

(14) Pan, J. Large Language Model for Molecular Chemistry. *Nat. Comput. Sci.* **2023**, *3*, 5.

(15) Yoshimori, A.; Chen, H.; Bajorath, J. Chemical Language Models for Applications in Medicinal Chemistry. *Future Med. Chem.* **2023**, *15* (2), 119.

(16) Wen, N.; Liu, G.; Zhang, J.; Zhang, R.; Fu, Y.; Han, X. A Fingerprints Based Molecular Property Prediction Method Using the BERT Model. *J. Cheminform.* **2022**, *14*, 71.

(17) Ross, J.; Belgodere, B.; Chenthamarakshan, V.; Padhi, I.; Mroueh, Y.; Das, P. Large-Scale Chemical Language Representations Capture Molecular Structure and Properties. *Nat. Mach Intell.* **2022**, *4*, 1256−1264.

(18) Flam-Shepherd, D.; Zhu, K.; Aspuru-Guzik, A. Language Models Can Learn Complex Molecular Distributions. *Nat. Commun.* **2022**, *13*, 3293−3303.

(19) Blanchard, A. E.; Gounley, J.; Bhowmik, D.; Chandra Shekar, M.; Lyngaas, I.; Gao, S.; Yin, J.; Tsaris, A.; Wang, F.; Glaser, J. Language Models for the Prediction of SARS-CoV-2 Inhibitors. *Int. J. High Perform Comput. Appl.* **2022**, *36*, 587−602.

(20) Jablonka, K. M.; Schwaller, P.; Ortega-Guerrero, A.; Smit, B. Is GPT-3 All You Need for Low-Data Discovery in Chemistry? *ChemRxiv Preprint*, 2023. https://chemrxiv.org/engage/chemrxiv/article-details/63eb5a669da0bc6b33e97a35 (accessed Feb 27, 2023).

(21) Rosset, C. *Turing-NLG: A 17-billion-parameter language model by Microsoft*, 2020. *Microsoft Research Blog*. https://www.microsoft.com/en-us/research/blog/turing-nlg-a-17-billion-parameter-language-model-by-microsoft/ (accessed February 19, 2023).

(22) Rae, J.; Irving, G.; Weidinger, L. *Language modelling at scale: Gopher, ethical considerations, and retrieval*. Deep Mind. https://www.deepmind.com/blog/language-modelling-at-scale-gopher-ethical-considerations-and-retrieval (accessed February 20, 2023).

(23) Brown, T. B.; Mann, B.; Ryder, N.; Subbiah, M.; Kaplan, J.; Dhariwal, P.; Neelakantan, A.; Shyam, P.; Sastry, G.; Askell, A.; Agarwal, S.; Herbert-Voss, A.; Krueger, G.; Henighan, T.; Child, R.; Ramesh, A.; Ziegler, D. M.; Wu, J.; Winter, C.; Hesse, C.; Chen, M.; Sigler, E.; Litwin, M.; Gray, S.; Chess, B.; Clark, J.; Berner, C.; McCandlish, S.; Radford, A.; Sutskever, I.; Amodei, D. Language Models Are Few-Shot Learners. *arXiv Preprint*, arXiv:2005.14165v4, 2020. https://arxiv.org/abs/2005.14165 (accessed Feb 27, 2023).

(24) Taylor, R.; Kardas, M.; Cucurull, G.; Scialom, T.; Hartshorn, A.; Saravia, E.; Poulton, A.; Kerkez, V.; Stojnic, R. Galactica: A Large Language Model for Science. *arXiv Preprint*, arXiv:2211.09085v1, 2022. https://arxiv.org/abs/2211.09085 (accessed Feb 27, 2023).

(25) Heaven, W. D. *Why Meta's latest large language model survived only three days online*, 2022. *MIT Technology Review*. https://www.technologyreview.com/2022/11/18/1063487/meta-large-language-model-ai-only-survived-three-days-gpt-3-science/ (accessed February 20, 2023).

(26) Shen, Y.; Heacock, L.; Elias, J.; Hentel, K. D.; Reig, B.; Shih, G.; Moy, L. ChatGPT and Other Large Language Models Are Double-edged Swords. *Radiology* **2023**, na DOI: 10.1148/radiol.230163.

(27) ChatGPT: Optimizing Language Models for Dialogue. *OpenAI*. https://openai.com/blog/chatgpt/ (accessed February 20, 2023).

(28) A simple guide to setting the GPT-3 temperature, 2020. *AlgoWriting*. https://algowriting.medium.com/gpt-3-temperature-setting-101-41200ff0d0be (accessed February 20, 2023).

(29) *What are tokens and how to count them? OpenAI*. https://help.openai.com/en/articles/4936856-what-are-tokens-and-how-to-count-them (accessed February 20, 2023).

(30) Weininger, D. SMILES, a Chemical Language and Information System: 1: Introduction to Methodology and Encoding Rules. *J. Chem. Inf. Comput. Sci.* **1988**, *28*, 31−36.

(31) White, A. D.; Hocky, G. M.; Gandhi, H. A.; Ansari, M.; Cox, S.; Wellawatte, G. P.; Sasmal, S.; Yang, Z.; Liu, K.; Singh, Y.; Peña Ccoa, W. J. Assessment of Chemistry Knowledge in Large Language Models That Generate Code. *Digit Discov.* **2023**, na DOI: 10.1039/D2DD00087C.

(32) Krenn, M.; Häse, F.; Nigam, A.; Friederich, P.; Aspuru-Guzik, A. Self-Referencing Embedded Strings (SELFIES): A 100% Robust Molecular String Representation. *Mach Learn Sci. Technol.* **2020**, *1*, 045024.

(33) Edwards, C.; Lai, T.; Ros, K.; Honke, G.; Cho, K.; Ji, H. Translation between Molecules and Natural Language. *arXiv Preprint*, arXiv:2204.11817v3, 2022. https://arxiv.org/abs/2204.11817 (accessed Feb 27, 2023).

(34) Griffin, S.; Wyllie, S. G.; Markham, J. Determination of Octanol−Water Partition Coefficient for Terpenoids Using Reversed-Phase High-Performance Liquid Chromatography. *J. Chromatogr A* **1999**, *864*, 221−228.

(35) Vilas-Boas, S. M.; da Costa, M. C.; Coutinho, J. A. P.; Ferreira, O.; Pinho, S. P. Octanol−Water Partition Coefficients and Aqueous Solubility Data of Monoterpenoids: Experimental, Modeling, and Environmental Distribution. *Ind. Eng. Chem. Res.* **2022**, *61*, 3154−3167.

(36) Costa, R. K. M.; Souza, L. M. P.; Silva, R. S.; Souza, F. R.; Pimentel, A. S. The Reconciliation between the Experimental and Calculated Octanol-Water Partition Coefficient of 1,2-Dipalmitoyl-

Sn-Glycero-3-Phosphatidylcholine Using Atomistic Molecular Dynamics: An Open Question. *J. Biomol Struct Dyn.* **2023**, 1−8.

(37) Souza, L. M. P.; Souza, F. R.; Reynaud, F.; Pimentel, A. S. Tuning the Hydrophobicity of a Coarse Grained Model of 1,2-Dipalmitoyl-Sn-Glycero-3-Phosphatidylcholine Using the Experimental Octanol-Water Partition Coefficient. *J. Mol. Liq.* **2020**, *319*, 114132.

(38) Fornasier, F.; Souza, L. M. P.; Souza, F. R.; Reynaud, F.; Pimentel, A. S. Lipophilicity of Coarse-Grained Cholesterol Models. *J. Chem. Inf Model.* **2020**, *60*, 569−577.

(39) Lancashire, R. J. Coordination Numbers and Geometry, 2020. *LibreTexts Chemistry*. https://chem.libretexts.org/Bookshelves/Inorganic_Chemistry/Supplemental_Modules_and_Websites_(Inorganic_Chemistry)/Coordination_Chemistry/Structure_and_Nomenclature_of_Coordination_Compounds/Coordination_Numbers_and_Geometry (accessed February 19, 2023).

(40) Kadajji, V. G.; Betageri, G. v. Water Soluble Polymers for Pharmaceutical Applications. *Polymers (Basel).* **2011**, *3*, 1972−2009.

(41) Symmetry and Point Groups. *Department of Chemistry and Physics, Arkansas State University*. http://myweb.astate.edu/mdraganj/pointgroupanswers.html (accessed February 19, 2023).

(42) Ross, J.; Belgodere, B.; Chenthamarakshan, V.; Padhi, I.; Mroueh, Y.; Das, P. Molformer: Large Scale Chemical Language Representations Capture Molecular Structure and Properties. *arXiv Preprint*, arXiv:2106.09553v3, 2022. https://arxiv.org/abs/2106.09553 (accessed Feb 27, 2023).

(43) Dunn, A.; Dagdelen, J.; Walker, N.; Lee, S.; Rosen, A. S.; Ceder, G.; Persson, K.; Jain, A. Structured Information Extraction from Complex Scientific Text with Fine-Tuned Large Language Models. *arXiv Preprint*, arXiv:2212.05238v1, 2022. https://arxiv.org/abs/2212.05238 (accessed Feb 27, 2023).

(44) Horawalavithana, S.; Ayton, E.; Sharma, S.; Howland, S.; Subramanian, M.; Vasquez, S.; Cosbey, R.; Glenski, M.; Volkova, S. Foundation Models of Scientific Knowledge for Chemistry: Opportunities, Challenges and Lessons Learned. In *Proceedings of BigScience Episode #5 - Workshop on Challenges & Perspectives in Creating Large Language Models*; Dublin, Ireland, May 27, 2022, Association for Computational Linguistics: Stroudsburg, PA, USA, 2022; pp 160−172.

https://doi.org/10.1021/acs.jcim.3c00285
J. Chem. Inf. Model. 2023, 63, 1649−1655