# Impact of noise on inverse design: The case of NMR spectra matching

## Supplementary Information

Dominik Lemm[1,2], Guido Falk von Rudorff[3,4] and O. Anatole von Lilienfeld[5,6,7]
[1] *University of Vienna, Faculty of Physics, Kolingasse 14-16, AT-1090 Vienna, Austria*
[2] *University of Vienna, Vienna Doctoral School in Physics, Boltzmanngasse 5, AT-1090 Vienna, Austria*
[3] *University Kassel, Department of Chemistry, Heinrich-Plett-Str.40, 34132 Kassel, Germany*
[4] *Center for Interdisciplinary Nanostructure Science and Technology (CINSaT), Heinrich-Plett-Straße 40, 34132 Kassel*
[5] *Departments of Chemistry, Materials Science and Engineering, and Physics, University of Toronto, St. George Campus, Toronto, ON, Canada*
[6] *Vector Institute for Artificial Intelligence, Toronto, ON, M5S 1M1, Canada*
[7] *Machine Learning Group, Technische Universität Berlin and Institute for the Foundations of Learning and Data, 10587 Berlin, Germany*
[*] Electronic address: anatole.vonlilienfeld@utoronto.ca

(Dated: 20 September 2023)

TABLE S1. Functional groups contained in the $C_7O_2H_{10}$ constitutional isomer chemical space and corresponding SMARTS patterns.

| **Functional Group** | **SMARTS Pattern** |
|---|---|
| alkene | [CX3]=[CX3] |
| alkyne | [CX2]#[CX2] |
| arene | [cX3]1[cX3][cX3][cX3][cX3][cX3]1 |
| alcohol | [#6][OX2H] |
| aldehyde | CX3H1[#6,H] |
| ketone | [#6]CX3[#6] |
| carboxylic acid | CX3[OX2H] |
| acid anhydride | CX3[OX2]CX3 |
| ester | [#6]CX3[OX2H0][#6] |
| ether | OD2[#6] |
| enol | [OX2H][#6X3]=[#6] |
| phenol | [OX2H][cX3]:[c] |

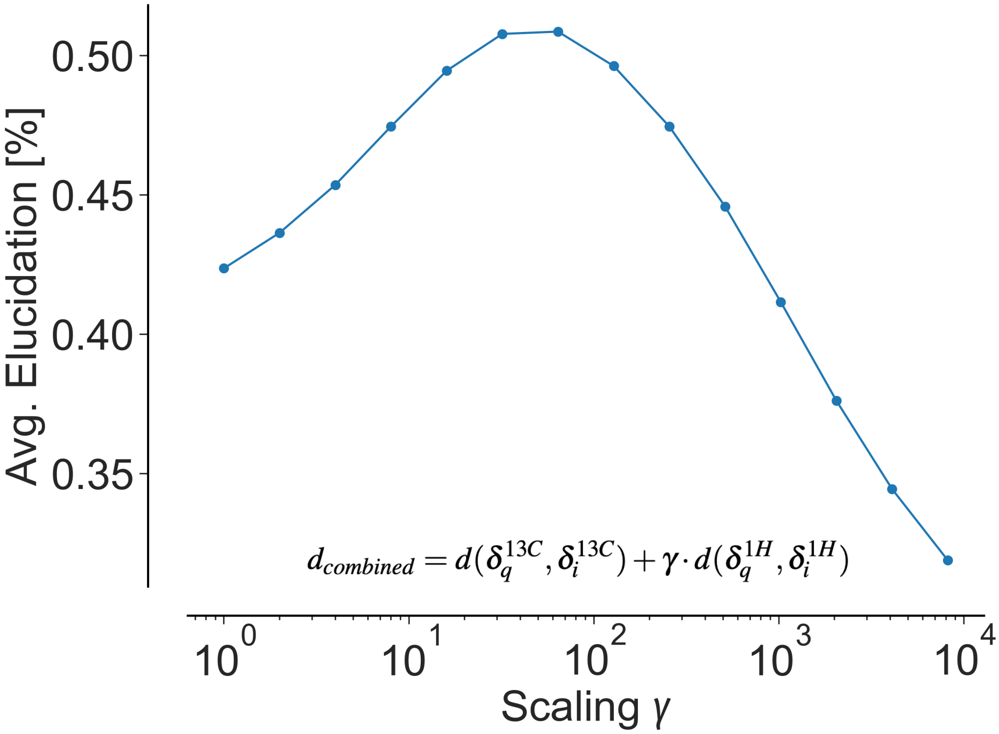

FIG. S1. Hyperparameter scan of $\gamma$ on $C_7O_2H_{10}$ constitutional isomers for the combined ranking of $^{13}$C and $^{1}$H shifts. First, the respective distances of $^{13}$C and $^{1}$H at their individual shift accuracy levels are being calculated and then the distances combined via the depicted Eq.2. The average elucidation is calculated by averaging across all shift accuracy levels.

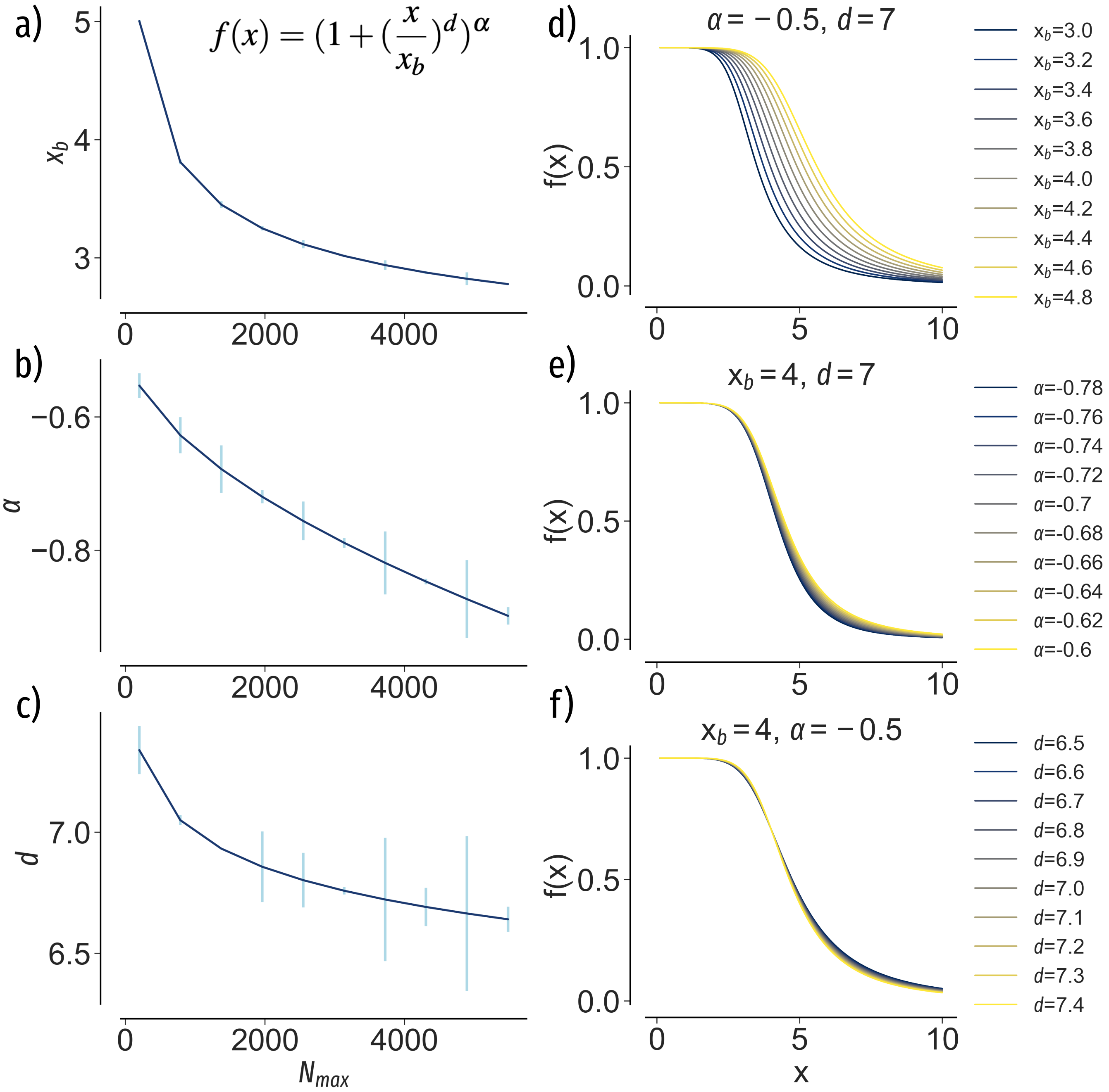

FIG. S2. Parameter distributions of a broken powerlaw function (Eq.3) used for extrapolating the elucidation trends. a-c) Parameters $x_b$, $\alpha$ and $d$ fitted to the elucidation trends of $C_7O_2H_{10}$ at multiple $N_{max}$. Note that the parameters $d$ and $\alpha$ are more noisy in nature given the finite sampling and only marginally influence the shape of the curve in the observed parameter range (see e) and f)). Conversely, the parameter $x_b$, which dictates the offset of the curve, is well behaved and decays smoothly as $N_{max}$ increases. d-f) Influence of the observed parameter ranges for $x_b$, $\alpha$ and $d$ on the shape of the broken powerlaw function.

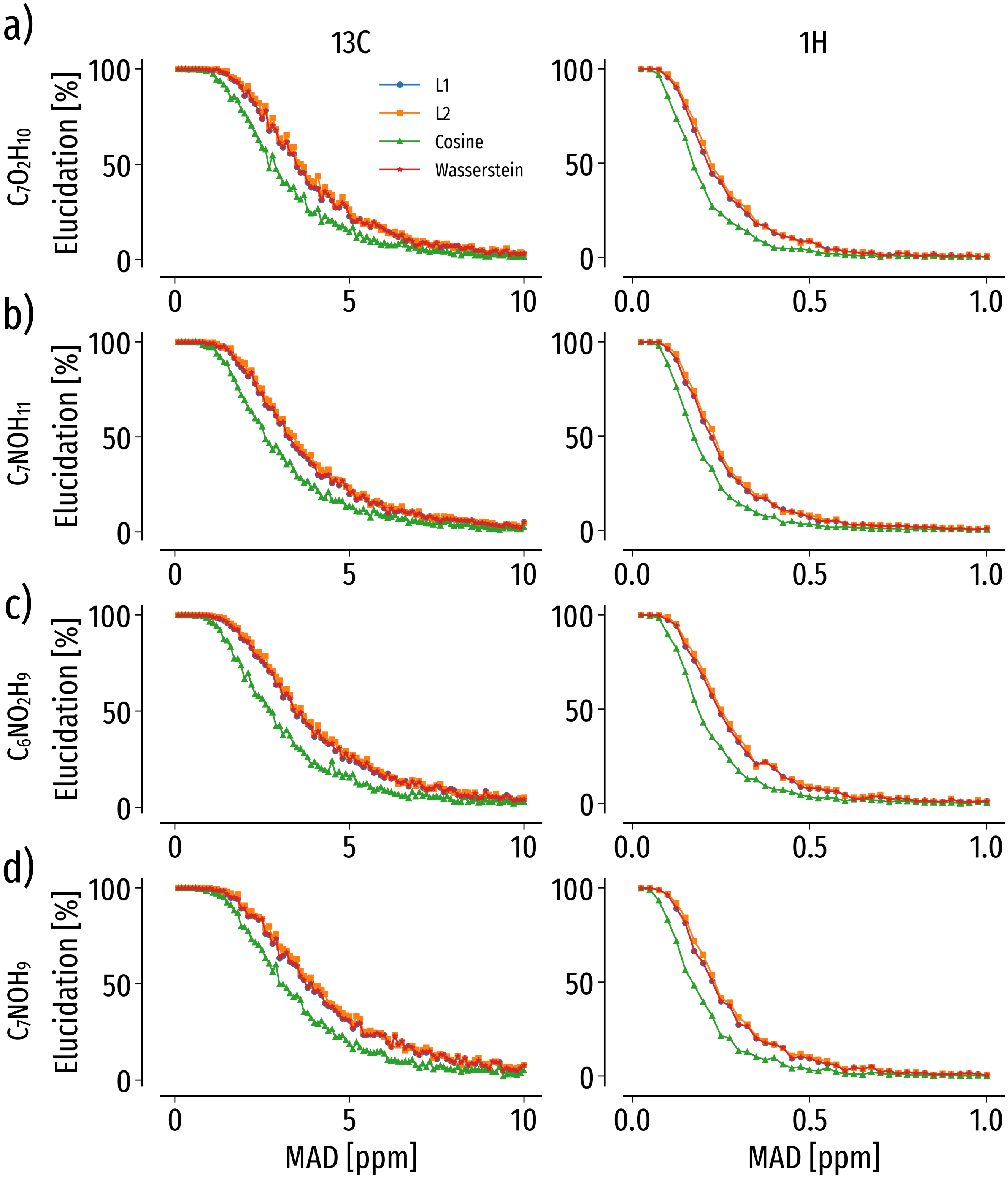

FIG. S3. Comparison of L1, L2, cosine similarity and Wasserstein distances on the $^{13}$C (left) or $^{1}$H (right) elucidation success of $C_7O_2H_{10}$ (a), $C_7NOH_{11}$ (b), $C_6NO_2H_9$ (c) and $C_7NOH_9$ (d) constitutional isomers.

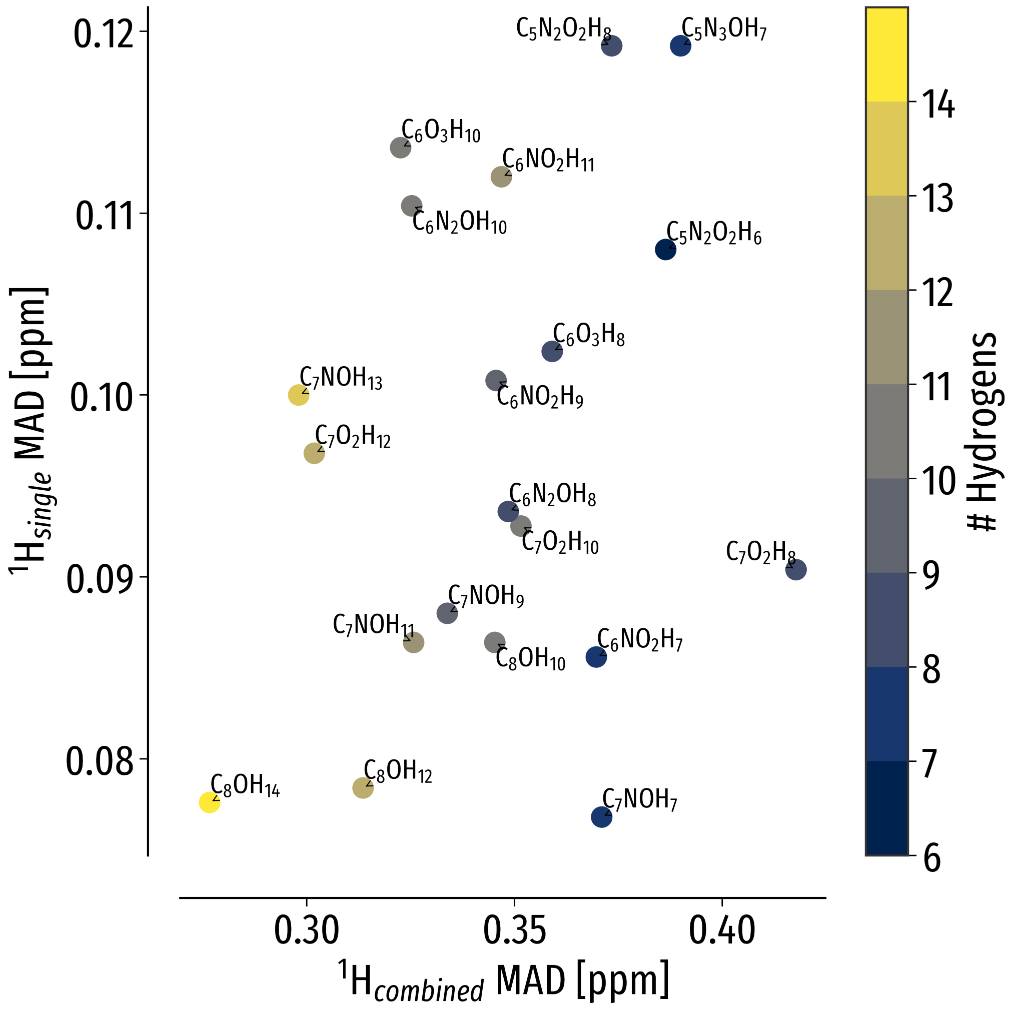

FIG. S4. Trends in QM9 chemical compound space to correctly elucidate queries at 95% accuracy. Mean absolute deviation (MAD) using only $^1$H spectra ($^1$H $_{single}$) against $^1$H and noise-free $^{13}$C spectra combined ($^1$H $_{combined}$) at the respective $N_{max}$ available in QM9.

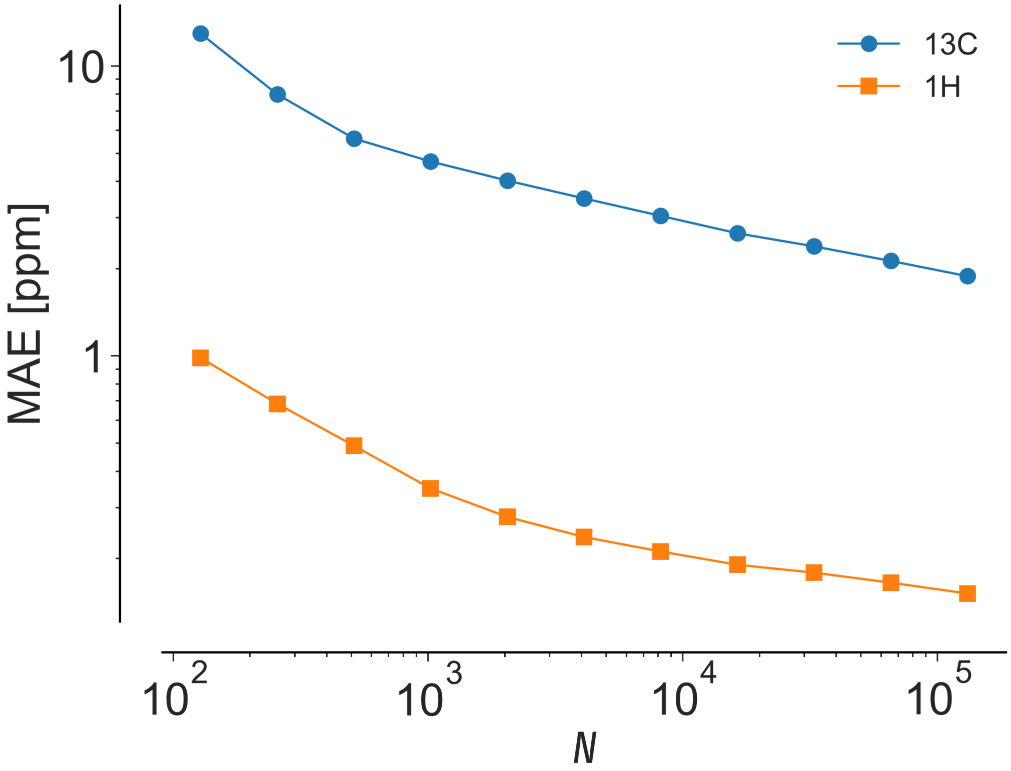

FIG. S5. Systematic improvement with increasing training set size *N* of KRR machine learning for $^{13}$C and $^{1}$H chemical shifts of $C_8OH_{12}$, $C_8OH_{10}$, $C_8OH_{14}$, $C_7O_2H_8$ and $C_7O_2H_{12}$ constitutional isomers using the FCHL19 representation.