

Supplementary Information

高质量的材料科学文本挖掘数据集构建方法*

刘悦¹⁾⁴⁾ 刘大晖¹⁾ 葛献远¹⁾ 杨正伟¹⁾ 马舒畅¹⁾ 邹喆义⁵⁾ 施思齐²⁾³⁾

1) (上海大学计算机工程与科学学院, 上海 200444)

2) (上海大学材料科学与工程学院, 上海 200444)

3) (上海大学材料基因组工程研究院, 上海 200444)

4) (上海市智能计算系统工程技术研究中心, 上海 200444)

5) (湘潭大学材料科学与工程学院, 湘潭 411105)

表 S1 NASICON 实体识别数据集的注释示例

Table S1. Examples of annotation for entity recognition dataset of NASICON.

实体类型	示例
Composition	The bond valence sum (BVS) of 1.08 is in good agreement with the value expected for Na+ .
Structure	The results from both measurements can be correlated only, if at room temperature Na+ is located in the interstitial sites and if the transition of these ions to the Na2 site occurs at high temperature.
Property	Conductivity measurements in the range 30 – 350 <sYm> reveal the activation energy of 0.3 eV for Na+ conduction but conductivity values were found to change with temperature of sample preparation.
Processing	Regrinding and reheating the mix results in very slow incorporation of the free ZrO ₂ , probably by replacement of Na+ from the zirconium sites.
Application	It is shown that Na ₃ TiP ₃ O ₉ N can reversibly cycle Na-ions in a manner suitable for secondary batteries , and that the volume changes on Na removal are remarkably small (<1%) with respect to other known Na-ion.
Characterization	The XRD diffraction pattern of Na ₃ MnTi(PO ₄) ₃ can be indexed into a rhombohedral NASICON type unit cell with the R3c space group.
Feature	Single crystal x-ray analysis is used to identify the composition NaZr ₂ P ₃ O ₁₂ and to refine its structure, which is rhombohedral.
Condition	The anisotropy of the thermal vibrations of sodium atoms in NaSn ₂ (PO ₄) ₃ at room temperature is described by two different flattened ellipsoids.

表 S2 NASICON 关系抽取数据集的注释示例

Table S2. Example of annotation for relational extraction dataset of NASICON.

关系类型	示例
Cause-Effect	<p>1. (<i>Composition-Property, Property-Property</i>) In the system $\text{Na}_{(1+x)}\text{Zr}_2\text{SixP}_{(3-x)}\text{O}_{12}$, on the other hand, the introduction of excess Na ions⁽¹⁾ causes electrostatic Na⁺-Na⁺ interactions⁽²⁾ that can lower the activation energy⁽³⁾ even though transport must be via a NaI site.</p> <p>2. (<i>Structure-Property</i>) The increase of the M1 site size⁽¹⁾ in $\text{Na}_2\text{SnFe}(\text{PO}_4)_3$ is accompanied by oxygen displacements perpendicular to the c axis which gives rise to rotation⁽²⁾ of the PO_4 tetrahedra and leads to a distortion⁽³⁾ of the $\text{Sn}(\text{Fe})_{(1-x)}(\text{PO}_4)_3$ framework.</p>
Component-Whole	<p>1. (<i>Composition-Composition</i>) The original Na super ionic conductors NASICON materials are solid solutions derived from NaZr2P3O12⁽¹⁾ by partial replacement of P⁽²⁾ by Si⁽³⁾ with extra Na⁽⁴⁾ to balance the charges.</p>
Feature-Of	<p>1. (<i>Composition-Property</i>) The calculated BVS value 5.22 shows that the As5 cation⁽¹⁾ is also slightly over bonded⁽²⁾.</p> <p>2. (<i>Composition-Structure</i>) The present paper reports on the crystal structure and vibrational spectra of NaZr2(AsO4)3.</p>
Located-Of	<p>1. (<i>Composition-Structure</i>) The Na cation⁽¹⁾ is located on the 6b position⁽²⁾ with a trigonal antiprismatic coordination and enhanced anisotropic displacement parameters.</p>
Instance-Of	<p>1. (<i>Property-Property</i>) The monochromator is a crystal of Ge that selects a wavelength⁽¹⁾ of 1.594 Å⁽²⁾.</p> <p>2. (<i>Feature-Feature</i>) X-ray powder diffraction shows that the phosphates⁽¹⁾ belong to the NZP type⁽²⁾.</p> <p>3. (<i>Structure-Structure</i>) The four P-O bonds⁽¹⁾ in the near regular tetrahedron⁽²⁾ (point symmetry₂) range from 1.524 to 1.525 Å, with O-P-O angles deviating by no more than 1.5 <sym> from the ideal 109.5 <sym> tetrahedral angle (Table III).</p>
Condition-On	<p>1. (<i>Processing-Condition</i>) For each crystal structure determination, data are collected using Mo Ka radiation⁽¹⁾ up to 29-650⁽²⁾.</p> <p>2. (<i>Property-Condition</i>) The disorder⁽¹⁾ is larger at 100 K⁽²⁾ than at 295 K⁽³⁾.</p>
Method-Of	<p>1. (<i>Characterization-Property</i>) The Rietveld plots⁽¹⁾ represent a good structure fit between observed and calculated intensity with satisfactory R-factors⁽²⁾.</p> <p>2. (<i>Characterization-Property</i>) Its anisotropic thermal expansion⁽¹⁾ has been calculated from high temperature X-ray diffraction⁽²⁾, and it is linear in a range from room temperature up to 800 <sym>.</p> <p>3. (<i>Processing-Condition</i>) Good crystals⁽¹⁾ can, however, be obtained after tempering⁽²⁾ in platinum crucible for several weeks at 11000 <sym>.</p>
Other	<p>1. (<i>Condition-Property</i>) At room temperature⁽¹⁾ no diffuse intensity⁽²⁾ is observed.</p>

表 S3 MatBERT-BiLSTM-CRF 模型参数

Table S3. Parameters of MatBERT-BiLSTM-CRF.

参数名称	值
批量大小 (Batch size)	32
迭代周期 (Epoch)	100
单词向量维度 (Word vector dimension)	768
LSTM 单元维度 (LSTM unit dimension)	128
丢包率 (Dropout rate)	0.1
学习率 (Learning rate)	3×10^{-5}
优化器 (Optimizer)	AdamW
提前停止耐力值 (Early stopping patience)	3
最大句子长度 (Max sentence length)	75

表 S4 不同机器学习模型的激活能预测结果^[33]

Table S4. Results of different machine learning models^[33].

Models	Dataset ₃₁			Dataset ₄₅		
	RMSE	MAPE	R^2	RMSE	MAPE	R^2
LASSO	0.09	0.06	0.86	0.06	0.04	0.94
GPR	0.09	0.06	0.86	0.05	0.04	0.96
Ridge	0.09	0.06	0.86	0.05	0.04	0.95
SVR	0.10	0.07	0.84	0.07	0.06	0.92
KNN	0.11	0.07	0.80	0.09	0.06	0.87
RF	0.10	0.06	0.83	0.05	0.04	0.96