



Original Paper

Predictive Geochemical Exploration: Inferential Generation of Modern Geochemical Data, Anomaly Detection and Application to Northern Manitoba

Julie E. Bourdeau ^{1,4}, Steven E. Zhang,¹ Christopher J. M. Lawley,¹ Mohammad Parsa,¹ Glen T. Nwaila,² and Yousef Ghorbani³

Received 21 April 2023; accepted 6 October 2023
Published online: 29 October 2023

Geochemical surveys contain an implicit data lifecycle or pipeline that consists of data generation (e.g., sampling and analysis), data management (e.g., quality assurance and control, curation, provisioning and stewardship) and data usage (e.g., mapping, modeling and hypothesis testing). The current integration of predictive analytics (e.g., artificial intelligence, machine learning, data modeling) into the geochemical survey data pipeline occurs almost entirely within the data usage stage. In this study, we predict elemental concentrations at the data generation stage and explore how predictive analytics can be integrated more thoroughly across the data lifecycle. Inferential data generation is used to modernize lake sediment geochemical data from northern Manitoba (Canada), with results and interpretations focused on elements that are included in the Canadian Critical Minerals list. The results are mapped, interpreted and used for downstream analysis through geochemical anomaly detection to locate further exploration targets. Our integration is novel because predictive modeling is integrated into the data generation and usage stages to increase the efficacy of geochemical surveys. The results further demonstrate how legacy geochemical data are a significant data asset that can be predictively modernized and used to support time-sensitive mineral exploration of critical minerals that were unanalyzed in original survey designs. In addition, this type of integration immediately creates the possibility of a new exploration framework, which we call predictive geochemical exploration. In effect, it eschews sequential, grid-based and fixed resolution sampling toward data-driven, multi-scale and more agile approaches. A key outcome is a natural categorization scheme of uncertainty associated with further survey or exploration targets, whether they are covered by existing training data in a spatial or multivariate sense or solely within the coverage of inferred secondary data. The uncertainty categorization creates an effective implementation pathway for future multi-scale exploration by focusing data generation activities to de-risk survey practices.

KEY WORDS: Machine learning, Geochemistry, Exploration, Manitoba, Critical Minerals.

¹Geological Survey of Canada, 601 Booth Street, Ottawa, Ontario K1A 0E8, Canada.

²Wits Mining Institute (WMI), University of the Witwatersrand, 1 Jan Smuts Avenue, Johannesburg 2000, Gauteng Province, South Africa.

³Joseph Banks Laboratories, School of Chemistry, University of Lincoln, Green Lane, Lincoln, Lincolnshire LN6 7DL, UK.

⁴To whom correspondence should be addressed; e-mail: Bourdeau.julie@nrcan-nrcan.gc.ca

INTRODUCTION

The desire to expedite mineral resource exploration to exploitation is evolving the mineral value chain through innovation and integration. There are two major axes of integration: (1) adoption of

transdisciplinary methods (artificial intelligence and data science) into the mineral industry (Karpatne et al., 2018; Bergen et al., 2019; He et al., 2022 and references therein) and (2) geometallurgical integration, of exploration, mining and mineral processing (e.g., Mena Silva et al., 2018). De-risked outcomes (e.g., survey data and prospectivity maps) are expected along both axes of integration and are an evolutionary pressure that is re-shaping the mineral value chain and its supporting disciplines (Mena Silva et al., 2018; Lawley et al., 2021; Zhang et al., 2021a, 2022b; Daviran et al., 2022; Ghorbani et al., 2022, 2023; Government of Canada, 2022; Shirmard et al., 2022). Geoscientific surveys contain three key stages of the data pipeline (or lifecycle): (1) data generation through sampling to analysis; (2) data management including data standardization, hosting and curation, quality assurance and control; and (3) data usage. The adoption of artificial intelligence almost entirely occurs during data usage, although its integration in earlier stages of the data pipeline and geoscience-specific methods are emerging research domains (He et al., 2022). A more thorough integration of transdisciplinary methods into exploration, particularly for the task of data generation, can increase survey agility by reducing survey time and cost and, hence, facilitate a timely discovery of mineral resources. Solutions to enhance exploration agility are becoming more pressing with time as traditional exploration practices intersect with emerging issues, such as resource depletion, sustainability, material criticality (e.g., critical raw materials, or CRMs) and supply chain security (Prior et al., 2012; Jowitt et al., 2018; Mudd, 2020; Michaux, 2021a, b).

Traditional geochemical surveys are variably scaled, statistically designed and are generally knowledge-driven, resulting in either systematic (e.g., grid-based) or local (e.g., deposit-scaled) datasets (Garrett 1983; Friske and Hornbrook, 1991; Grunsky, 2010; Demetrides et al., 2018). Geochemical data generated through surveys support mineral exploration. Traditional surveys exhibit several consistent characteristics: (1) approximately uniform spatial resolution; (2) systematic or grid-based execution; and (3) pervasive and substantial use of manual labor (Garrett 1983; Govett, 1983; Friske and Hornbrook, 1991; Friske, 1991; Demetrides et al., 2018). Over time, an accumulation of survey data gradually increases the proportion of brown-field to greenfield settings and results in a collection of legacy data that are no longer the state-of-the-art.

For example, legacy geochemical data from the Geological Survey of Canada (GSC) exclude analyses of rare earth elements (REEs), lithium, vanadium, beryllium and other CRMs. Consequently, as survey methods evolve, such as changes in instrumentation, legacy data are cyclically modernized through re-sampling and re-analysis (e.g., McCurdy et al., 2016; Council for Geoscience, 2022). Because the evolution of geochemical data has been guided by scientific reduction, it is expensive, and thus, large survey programs incur sizable continuous expenditures to repetitively update or acquire modern data. The high cost and low flexibility of geochemical surveys are promoting the use of cheaper, more agile and data-abundant reconnaissance methods to guide more targeted geochemical exploration (Sabins, 1999; McCaffrey et al., 2005; Booyesen et al., 2019; Shirmard et al., 2022).

Geochemical data (and other types of geoscientific data) are used with artificial intelligence algorithms to automate data modeling, which requires the formulation of geoscientific tasks into artificial intelligence or more specifically machine learning tasks. Common machine learning tasks in mineral exploration can be categorized into: (1) data analysis (Lary et al., 2016; Alférez et al., 2022; He et al., 2022 and references therein); (2) prospectivity mapping (e.g., Rodriguez-Galiano et al., 2015; Lawley et al., 2021); and (3) data inversion (e.g., Kirkwood et al., 2016; Zhang et al., 2023). In data analysis, machine learning algorithms replace manual data modeling and/or knowledge-driven methods. In prospectivity mapping, machine learning algorithms model relationships between high-dimensional and multidisciplinary evidence layers and known deposit locations to predict the prospectivity of unknown locations. In data inversion, because geochemical data are expensive but necessary for a range of tasks, other types of data are inverted into geochemical data using machine learning models. Supervised machine learning is used in (1) to (3), while unsupervised methods are usually used in (1) and (2), because (3) requires data labels. For the purpose of exploration, machine learning methods are intended to narrow the search space for various deposits. This is the premise of prospectivity mapping (e.g., Rodriguez-Galiano et al., 2015; Lawley et al., 2021; Zhang et al., 2021a). Unsupervised methods make use of clustering (e.g., Zhang et al., 2021c), dimensionality reduction (e.g., Grunsky and de Caritat, 2020) and anomaly detection (He et al., 2022 and references therein) to segment the data to

aid analysis. Where data are of suitable abundance and quality, deep learning can be used, for example, to identify prospective areas and to extract geologically relevant geochemical anomalies (e.g., Zhang et al., 2022a; Zuo et al., 2022; Zuo and Xu, 2023). Extracted anomalies constitute a type of targeting insight that could be used to guide further activities (e.g., reconnaissance or exploration). There are some general issues known in the formulation of geoscientific tasks that include a reliance on repurposing of reductive scientific data, which is expensive and unnecessary for big data methods (e.g., He et al., 2022 and references therein, also Rodriguez-Galiano et al., 2015; Sun et al., 2019; Lawley et al., 2021, 2022; Zhang et al., 2021a, 2022b). Some other issues are specific to the nature of geosciences, such as algorithmic considerations for spatial variability and model explainability (Karpatne et al., 2018; Hoffmann et al., 2021; He et al., 2022 and references therein). Model explainability is of serious concern for scientists, because the tradition of science is to create simple, elegant models coupled with reliable knowledge. Unfortunately, this aspect becomes challenging for more complex models, such as models resulting from the use of deep learning (Zuo et al., 2019; Linardatos et al., 2021).

Despite exploration activities becoming routinely formulated into artificial intelligence tasks, there are no documented approaches to transform core survey practices, which are data generation. All known data analysis must necessarily occur downstream of data generation. Weaknesses of geochemical data generation through grid-based surveys include a diminishing return in brownfield settings and a lack of agility to pursue time-sensitive exploration, resulting in geochemical data not being used for critical exploration activities (e.g., Lawley et al., 2021, 2022). One solution to better focus geochemical data generation is to predict ahead of geochemical re-sampling or re-analysis campaigns, because there already exist legacy data in brownfield settings, which could be inferentially modernized (Zhang et al., 2022b). Availability of inferred data implies that subsequent activities could be inferentially informed. For this purpose, Zhang et al. (2022b) formulated geochemical re-analysis campaigns into a machine learning task that leverages existing survey data to conduct essentially zero-cost geochemical reconnaissance in brownfield settings, by noticing that because modern and legacy data share samples, it is possible to predict modern from

legacy data without incurring additional analyses or sampling. This is a purely geochemical reconnaissance, unlike data inversion (e.g., Zhang et al., 2023) and is very different than data analysis using artificial intelligence, because the latter combination yields information from existing data, whereas the former increases the number of elements analyzed inferentially. This approach realizes the value of legacy data, which are not routinely used anymore due to availability of modern data (through re-analysis or surveys), and the cyclic nature of survey programs.

In this paper, we deploy the method of inferential data generation in an exploration pipeline, by predicting (inferring) modern data from legacy data. Furthermore, we couple the generated data with anomaly detection. The main novelty of our dual-use approach lies in the demonstration of a new proposed data-driven exploration framework. We produce the first (inferred) modern data of the northern Manitoba region using GSC's geochemical datasets, without incurring any additional survey costs. We focus on geochemistry because its data yield the most direct information regarding CRMs, which are exploration time-sensitive. We present our inferred elemental concentration maps and anomalies for select elements, focusing on unanalyzed elements in legacy data, which include nickel, rare earth elements (REEs), lithium and tungsten. Because geochemical anomaly detection is no longer novel, our usage of it is to fulfill the minimum requirements of our proposed framework. We imagine that our framework could be used to transform geochemical survey programs into a predictor-corrector operation, away from a systematic, grid-based operation. We term our framework 'predictive geochemical exploration.' It is important to distinguish our conception of predictive geochemical exploration with geochemical exploration or prospecting. Traditional exploration and prospecting generate primary (actual) geochemical data using resources, whereas our framework generates inferred geochemical data using existing data, which are then analyzed to steer primary data generation. Therefore, the expected use of our framework is not exploration targeting in the traditional sense (e.g., mineralization targets), but is to infer unknown elemental distributions to guide core survey activities (e.g., sampling or re-analysis). Anticipated benefits of predictive geochemical exploration include: (1) smaller and more targeted survey designs; (2) timely and more focused prospectivity

maps; and (3) data-driven focusing of exploration expenditures.

GEOLOGICAL CONTEXT

Northern Manitoba is part of the Canadian Shield, which is a large area of exposed Precambrian igneous and high-grade metamorphic rocks. The data used in this study cover areas that encompass the southwestern portion of the Churchill Province and the northwestern portion of the Trans-Hudson Orogen (Fig. 1).

The present study is focused on the southern limit of the Neoproterozoic Hearne Craton, which is one of the largest Neoproterozoic 'greenstone' terranes in the Canadian Shield (Hoffman, 1988). The oldest rocks correspond to juvenile submarine to subaerial volcanic assemblages that were formed at 2.71–2.69 Ga (Davis et al., 2004; Hanmer et al., 2004). Volcanic rocks are predominantly tholeiitic and calc-alkaline basalts with lesser proportions of intermediate to felsic rocks. At 2.69–2.68 Ga (Davis et al., 2004), the Hearne Craton volcanic assemblages were deformed, metamorphosed (greenschist-facies), intruded by calc-alkaline plutons and covered by turbidite-dominated sedimentary rocks. Afterward, post-deformation granites were emplaced at 2.67 Ga, followed by the deposition of polymict conglomerates and the intrusion of alkaline magmas, including carbonatites (< 2.66 Ga; Davis et al., 2004; Hanmer et al., 2004). The mineral potential of the southern Hearne Craton has never been fully investigated due to inaccessibility, lack of exposure (i.e., coverage of the area by glacial till) and lack of geoscientific data (Davies et al., 1962). The few known occurrences and deposits comprise anomalous Fe, U, Pb, Zn and REE concentrations from hand samples and other geochemical anomalies (Manitoba Mineral Resources, 2013; Saskatchewan Geological Survey, 2018; Tschirhart et al., 2022).

Paleoproterozoic arc terranes separate the southern Hearne Craton from the Superior Craton to the west (Fig. 2). These metamorphosed and deformed volcanic arcs formed during a series of accretionary events that occurred prior to the final closure of the Manikewan Ocean (Stauffer, 1984; Corrigan et al., 2005, 2007, 2009). The subsequent collision (Trans-Hudson Orogen) between the Hearne and Superior cratons at 1.84–1.83 Ga deformed and metamorphosed the intervening ac-

creted arc terranes (Corrigan et al., 2007, 2009). The western portion of the Trans-Hudson Orogen is divided into the: (1) Southern Indian, (2) Lynn Lake, (3) Flin Flon and (4) Kisseynew domains, together with the 1.86–1.85 Ga Chipewyan-Wathaman batholith (Fig. 2; MacHattie, 2001; Martins et al., 2021). Numerous mineral systems are associated with the complex accretionary and collisional history of the Trans-Hudson Orogen in this part of northern Manitoba. For example, the juvenile volcanic arcs in the Flin Flon, Lynn Lake and Leaf Rapids domains are associated with numerous volcanogenic massive sulfide (VMS; Cu, Zn, Pb, Ag and Au) deposits that formed as part of accretionary orogenesis (e.g., Lallor mine; Fox deposit). Magmatic (Ni, Cu and PGE) deposits also formed during accretionary orogenesis, possibly due to back-arc rifting and associated arc magmatism (e.g., Lynn Lake deposit). Sedimentary rocks comprising the Kisseynew Basin were deposited during this type of rift event and prior to the final collision between the Hearne and Superior cratons. Pegmatites that intrude the Kisseynew Basin and the younger volcanic arcs are the product of collisional orogenesis and are likely important hosts for critical minerals. Finally, orogenic (Au) deposits are associated with all accretionary and collisional stages, including lesser-known metamorphic rocks comprising the Southern Indian domain.

During the Wisconsin glaciation, northern Manitoba was covered by the continental-scale Laurentide Ice Sheet, which profoundly eroded and modified the Precambrian surface. The glaciation remobilized surface material in a general southwest direction (Dredge and McMartin, 2011). Subsequent melting of the ice sheet resulted in the deposition of large amounts of glacial drift material. As of current day, drainage within Manitoba is toward the northeast (Hudson Bay), thus, in the context of lake sediment sampling, counteracting some degree of glacial material remobilization.

DATA DESCRIPTION AND METHODS

Lake Sediment Samples, Analysis and Data

The data cover both deployment and training regions (Figs. 1 and 3). The deployment area contains mostly legacy geochemical analyses of 13,056 lake sediment samples (Fig. 3a; Table 1). These samples were collected from 1976 to 1991 by the GSC as part of the national uranium reconnaissance

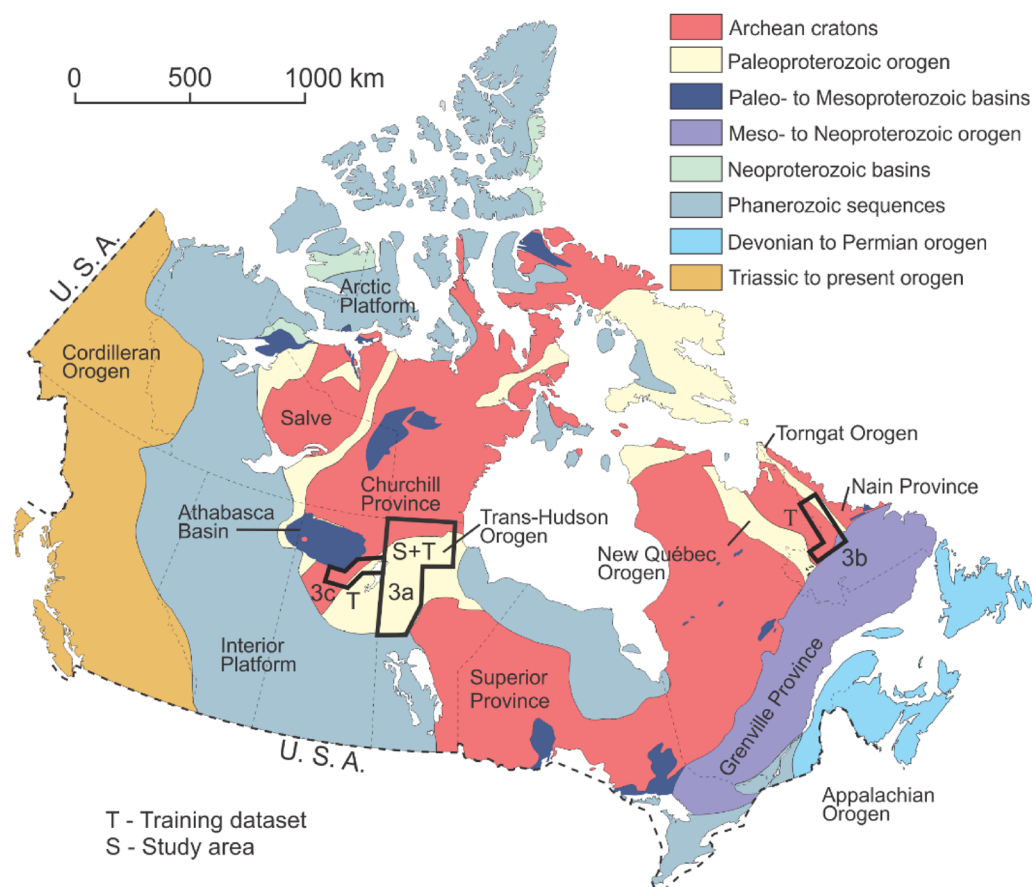


Figure 1. Simplified geological map of Canada with the location of the training datasets (T) and of the study or deployment area (S) in Manitoba, Labrador and Saskatchewan. Detailed sample distributions, here labeled 3a, b and c are presented in Figure 3a, b and c, respectively.

(1975-1976), Manitoba mineral development agreement (1984-1989) and the exploration science and technology initiative (1991) programs. Sample is located in NTS zones 063 K, J, N and O, 064B, C, F, G, I, J, K, N, O and P, and 054L and M. A subset of these legacy samples that are located in NTS zones 064F, I, K, N, O and P, and 54L and M were re-analyzed in 2010 and 2022 (Fig. 3a) and are used for training as part of the current study. The training data from the Labrador region contain a total of 3,441 samples and were re-analyzed in 2016 (Fig. 3b). The training data from the northeastern Saskatchewan region contain a total of 2,970 lake sediment samples and were re-analyzed in 2020 (Fig. 3c). The Saskatchewan and Labrador samples are located in NTS zones 074A, B, G and H, 064E, 014D, 013M and L, and 023I and J.

Lake sediment samples are taken from the center of medium-size lakes, preferably measuring

between 1-5 km in length and ≥ 3 m in depth (Friske and Hornbrook, 1991; Friske, 1991; Cameron, 1994; Bourdeau and Dyer, 2023). Samples are collected at an approximate density of 1 per 13 km² (or ~ 5 mi²). Collected samples were air dried and subsequently crushed and milled to the - 80 mesh (177 μ m) prior to geochemical analyses. Depending on the specific dataset, legacy geochemical analyses were performed by Chemex Laboratories (now ALS Global), Barringer Magenta Ltd. and Becquerel Laboratories from 1975 to 1991. Most elements were determined using atomic absorption spectroscopy (AAS), with U determined via instrumental neutron activation analysis (INAA) (e.g., Table 2). Typical legacy analyses contain a minimum of 13 elemental concentrations (Table 1). Modern geochemical datasets contain a maximum of 65 elemental analyses and were performed by Activation Laboratories Limited (2010) and Bureau Veritas (2022). Pulped

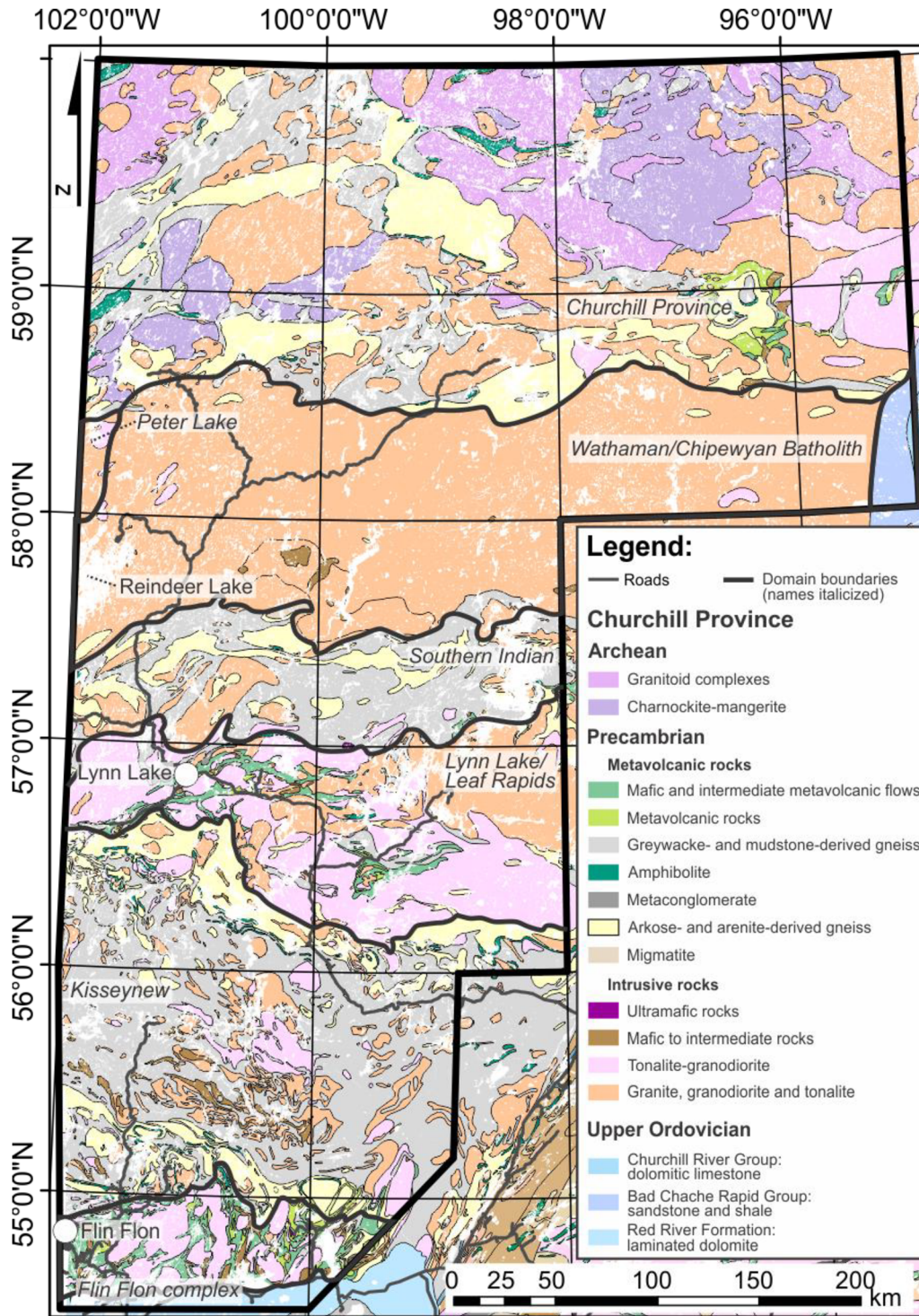


Figure 2. Geological map of the northern Manitoba region, which is the deployment region of machine learning models. Geological map obtained and modified from Manitoba Mineral Resources (2013). Outlines between geological domains are shown, and an outer polygon (shown as a thick black line) depicts the extent of legacy sample coverage in the area.

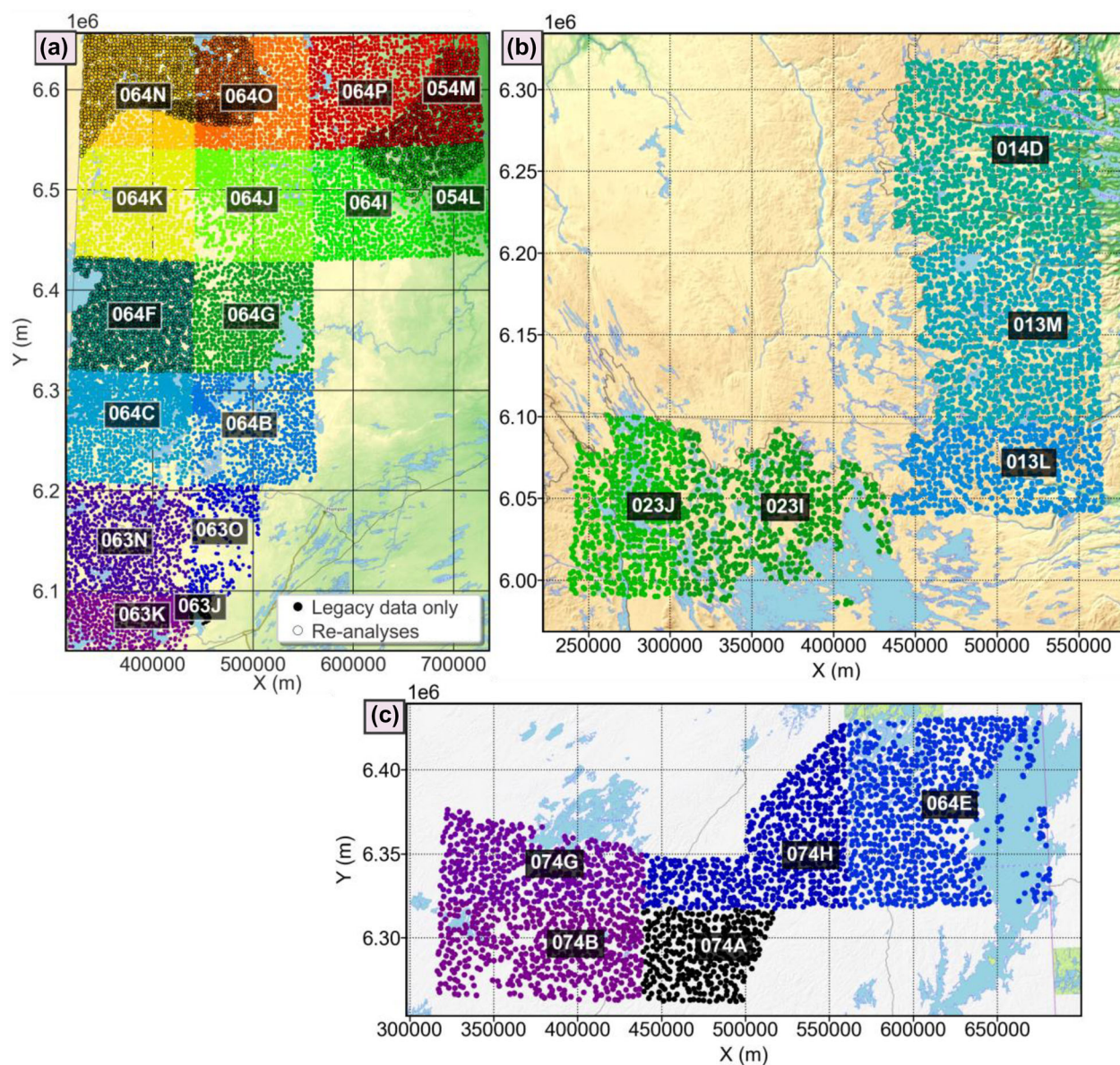


Figure 3. Location of lake sediment samples in the deployment region in northern Manitoba (a), training data coverage in Labrador (b) and Saskatchewan (c). The locations of the three areas within Canada are presented in Figure 1. Each color corresponds to an NTS zone. The deployment region in Manitoba (a) strictly excludes modernized data (re-analyses), which were used for training instead. Modernized data coverage in Manitoba is variable. Rare earth elements Pr to Lu have not been re-analyzed in the northeastern portion (NTS zones 054M, L and 064I, P). Element Hg has not been re-analyzed in the northwestern portion (NTS zones 064K, N and O). Both Labrador and Saskatchewan datasets have been completely re-analyzed and were used for model training.

samples were digested using a mixture of $\text{HCl}:\text{HNO}_3:\text{H}_2\text{O}$ (1:1:1). Digested samples were determined via ICP-MS or ICP-emission spectrometry (ICP-ES). Quality assurance and quality control (QA/QC) was achieved in re-analyzed datasets by including reference materials, analytical and field duplicate samples (McCurdy and Garrett, 2016).

The published legacy datasets did not contain analyses of reference materials, which implies that their accuracy could not be determined. However, the precision of the legacy and modern datasets could be determined and are presented in Table 2. Both legacy and modern geochemical data were fused on a

Table 1. Details of Manitoba geochemical datasets used in this study

Zone	Year	Open File	Samples	Legacy Open File	Elemental concentrations determined
64N	1975	322	1068		Ag, As, Co, Cu, Fe, Hg, Mn, Mo, Ni, Pb, U, Zn, LOI
64O	1975	323	1077		
64P, 54M	1976	408	1368		
64K	1975	321	1012		
64J	1975	320	993		
64I, 54L	1976	407	1387		
64F	1984	1104	966		Ag, As, Cd, Co, Cu, Fe, Hg, Mn, Mo, Ni, Pb, Sb, U, V, Zn, LOI
64G	1984	1105	881		
64C	1983	1288	1293		Ag, As, Au, Cd, Co, Cu, Fe, Hg, Mn, Mo, Ni, Pb, Sb, U, V, Zn, F, LOI
64B	1984	1287	747		
64B, C	1985-1988	1959	394		Ag, As, Au, Cd, Ba, Co, Cu, Fe, Hg, Mn, Mo, Ni, Pb, Sb, Sn, U, V, W, Zn, F, LOI
63N, O	1985	1212	1524		Ag, As, Au, Cd, Co, Cu, Fe, Hg, Mn, Mo, Ni, Pb, Sb, U, V, Zn, F, LOI
63K, J	1991	3015	346		Ag, Cd, Co, Cu, Fe, Hg, Mn, Mo, Ni, Pb, V, Zn, F, LOI
	2010	7309	737	322	Ag, Al, As, Au, B, Ba, Be, Bi, Ca, Cd, Ce, Co, Cr, Cs, Cu, Dy, Er, Eu, Fe, Ga, Gd, Ge, Hf, Ho, In, K, La, Li, Lu, Mg, Mn, Mo, Na, Nb, Nd, Ni, P, Pb, Pr, Rb, Re, S, Sb, Sc, Se, Sm, Sn, Sr, Ta, Tb, Te, Th, Ti, Tl, Tm, U, V, W, Y, Yb, Zn, Zr
	2010	7309	456	323	
	2010	6671	501	408	
	2010	7309	5	321	
	2010	6771	305	407	Ag, Al, As, Au, B, Ba, Be, Bi, Ca, Cd, Ce, Co, Cr, Cs, Cu, Fe, Ga, Ge, Hf, Hg, In, K, La, Li, Mg, Mn, Mo, Na, Nb, Ni, P, Pb, Pd, Pt, Rb, Re, S, Sb, Sc, Se, Sn, Sr, Ta, Te, Th, Ti, Tl, U, V, W, Y, Zn, Zr
	2022	9848	966	1104	Ag, Al, As, Au, B, Ba, Be, Bi, Ca, Cd, Ce, Co, Cr, Cs, Cu, Dy, Er, Eu, Fe, Ga, Gd, Ge, Hf, Hg, Ho, In, K, La, Li, Lu, Mg, Mn, Mo, Na, Nb, Nd, Ni, P, Pb, Pd, Pr, Pt, Rb, Re, S, Sb, Sc, Se, Sm, Sn, Sr, Ta, Tb, Te, Th, Ti, Tl, Tm, U, V, W, Y, Yb, Zn, Zr

LOI Loss on ignition

per-sample basis using the sample unique identification to construct a dataset for predictive modeling.

Machine Learning Workflows

There are two machine learning tasks in our study (e.g., two workflows; Fig. 4). The first task is to modernize legacy geochemical data (inferential data generation). The second task is geochemical anomaly detection to determine prospective locations for further exploration.

For inferential data generation, we use the exact workflow that was published by Zhang et al. (2022b), which evolved from previous machine learning-based workflows to predict trace element

concentrations (Zhang et al. 2021a, b). Inferential data generation relies on two facts (Zhang et al., 2022b): (1) survey data are cyclically modernized following obsolescence of legacy data; and (2) integrating legacy and modern data provides training data for machine learning algorithms, such that trained models could be deployed to regions containing solely legacy data. This workflow roughly followed a standard machine learning workflow that contains: (1) exploratory analysis and data pre-processing; (2) predictive modeling; and (3) mapping and post-hoc assessments (Fig. 4). Key subtasks in data pre-processing included: matching of legacy and modern data records using a unique key; data consistency checks and measures (e.g., standardization of units); and data imputation. Data leveling of

Table 2. Comparison of precision between legacy (1984) and modern (2022) geochemical data for zone NTS 064F (fully re-analyzed) in northern Manitoba (Fig. 3)

Element	1984 data via AAS and U via INAA				2022 re-analysis data via ICP-MS			
	Mean (all data)	STD (all data)	Percentage censored LDL	RSD (%)	Mean (all data)	STD (all data)	Percentage censored LDL	RSD (%)
Ag			100		0.06	0.02	0	6.16
As	1.89	1.32	84.48	0.00	0.80	0.76	8.62	13.16
Cd	0.20	0.00	86.21	0.00	0.47	0.13	0	4.07
Co	8.11	5.27	1.72	8.65	8.77	5.27	0	4.40
Cu	14.02	5.49	0	5.30	13.49	6.05	0	4.93
Fe (%)	4.39	4.61	0	2.64	4.10	4.15	0	2.20
Hg	0.06	0.02	0	12.74	0.08	0.03	0	13.27
Mn	502.93	622.04	0	2.03	480.93	599.57	0	2.83
Mo	3.23	2.54	55.17	21.03	2.56	2.39	0	5.25
Ni	11.44	5.52	0	9.97	13.96	6.44	0	3.05
Pb	2.88	1.26	86.21	12.30	3.84	1.81	0	3.26
Sb			100		0.06	0.03	1.72	14.54
U	3.18	2.84	1.72	12.32	2.74	2.93	0	3.56
V	31.18	18.89	1.72	11.14	31.16	18.17	0	1.88
Zn	114.33	42.49	0	3.59	114.89	36.47	0	3.22

For the calculation of RSD, both mean and standard deviation are calculated between pairs of samples and are detailed in McCurdy and Garrett (2016). All elemental arithmetic means, unless otherwise stated, are given in ppm
STD standard deviation, *LDL* lower detection limit, *RSD* relative standard deviation

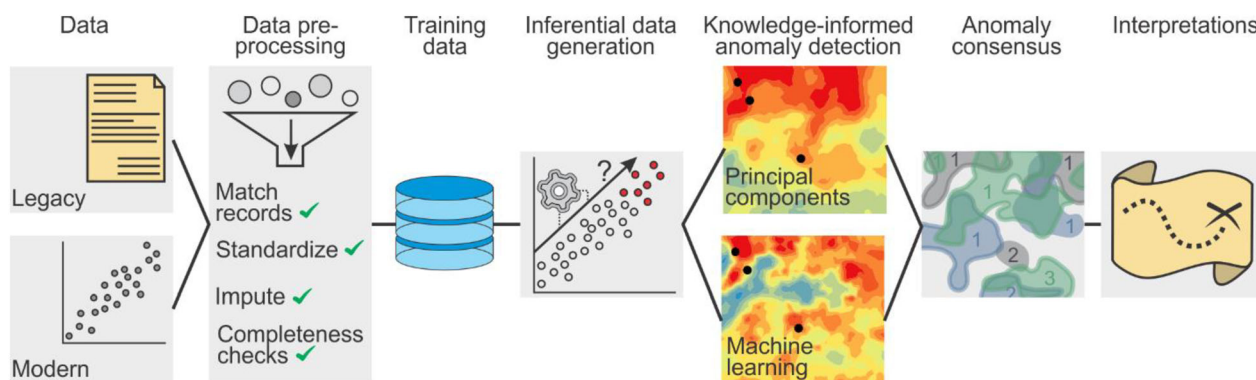


Figure 4. Schematic workflow used in this study. The order for tasks begins to the left and progresses to the right.

legacy data was not feasible, and in a similar previous task, it was observed that data generation performance was sufficient using even significantly unlevelable (secondary) data (Zhang et al., 2021b, 2022b). During predictive modeling, within the discipline of data science, there is no practical limit to the diversity of algorithms that could be tested. A variety of algorithms were indeed explored, and some were consistently better in previous method development studies of a similar purpose (e.g., Zhang et al., 2021a, b). Among shallow learning algorithms, random forest is generally better than

other algorithms for the task of inferential data generation (Zhang et al., 2021a, b). Following the demonstration of inferential data generation, the focal novelty had shifted toward its practical usage in solely data generation (Zhang et al., 2022b). In this study, the focus is around a proof-of-concept deployment of the inferential data generation method into a new area without prior ground truth. Consequently, we chose to de-emphasize the data science aspect of the exploration of multiple machine learning algorithms, which was well explored in (Zhang et al., 2021a, b). In this study and fol-

lowing (Zhang et al., 2022b), we focus on solely the random forest algorithm and demonstrate the effectiveness of inferential data generation in a deployment scenario.

For geochemical anomaly detection, the workflow was established by Zhang et al. (2021a, b), which used major and minor elemental concentrations to re-construct trace elemental concentrations (Fig. 4). The reconstruction error (prediction residual) was used as a proxy to geochemical anomalies (predicted minus actual concentrations in Zhang et al., 2021a, b; in this study, adopting the convention of traditional geochemical data analysis, we use actual minus predicted). The workflow was similar to that of the inferential data generation task and included: (1) data pre-processing; (2) predictive modeling; and (3) mapping and post-hoc assessments. Similar to that of the previous task, we forgo algorithm exploration and focus on the tuning of the random forest algorithm, because although reconstruction performance generally differs quantitatively between various algorithms, the anomaly maps are similar and the random forest algorithm is a top performer in general across a range of elements for precisely our machine learning task (Zhang et al. 2021a, b).

Both workflows make use of two machine learning algorithms, random forest for predictive modeling and k-nearest neighbors for imputation of only features as part of data pre-processing. Random forest is a type of ensemble algorithm that averages the output of a collection of decision trees. Decision trees are hierarchical flowchart-type structures that use nodes to represent features, branches to represent decision rules and leaves to represent outcomes. Decision trees learn to partition data based on feature values. The decision to split a node into finer sub-nodes is metric-driven to minimize model error. To produce a statistically meaningful average, the random forest algorithm constructs an ensemble of de-correlated decision trees (bagging) (Ho, 1995; Breiman, 1996a, b; Kotsiantis, 2014; Freund and Schapire, 1997; Sagi and Rokach, 2018). De-correlation of decision trees occurs via bootstrap sampling of features for each tree. Bagging lowers the noise sensitivity of decision trees and allows a more accurate model output by lowering the model variance without incurring additional bias. Hyperparameters of the random forest algorithm in addition to those in the decision tree algorithm include: maximum number of features per tree, the number of trees, tree depth and the mini-

um number of samples per split. In geochemical data generation, properties of the random forest algorithm, such as the lack of native feature space geometry, makes low demands on the pre-processing of data. In this case, it is not theoretically or empirically necessary to perform an embedding of the geochemical data through a transform (e.g., log-ratio transforms, see a thorough exploration in Zhang et al., 2021a, b). The k-nearest neighbors algorithm (Cover and Hart, 1967; Fix and Hodges, 1951) is a nonparametric method that uses an average of the labels of the closest training samples in feature space to derive an estimate for a target (Kotsiantis, 2014; Witten and Frank, 2005).

The features used for both machine learning tasks were different. For inferential data generation, elements that were analyzed in the legacy datasets (prior to 2010) that were nearly complete (> 90%) were used as features (Co, Cu, Fe, Hg, Ni, U and Zn), which is a machine learning vocabulary that refers to predictor variables. The other elements in the modern analyses were used as targets (Ag, Al, Ba, Be, Ca, Cd, Ce, Co, Cr, Cs, Cu, Dy, Er, Eu, Fe, Ga, Gd, Hg, Ho, K, La, Li, Lu, Mg, Mn, Mo, Na, Nb, Nd, Ni, P, Pb, Pr, Rb, S, Sc, Se, Sm, Sr, Tb, Th, Ti, Tl, Tm, U, V, Y, Yb, Zn and Zr). All elemental concentrations that were more than 50% uncensored were used as training data in inferential data generation (elements not meeting this criterion were B, Ge, Hf, In, Pd, Pt, Re and Ta; see Fig. 5). Censoring of data labels in the targets was not imputed, and hence, all censored samples on a per-target element basis were discarded for model training. Censoring of data labels in the features were imputed using a k-nearest imputer similar to that in Zhang et al. (2021a, b, 2022b). For anomaly detection, the features were all the major and minor elements (Al, Ca, Fe, K, Mg, Na, P, S and Ti) and the targets were all of the remainder elements. Hence, we used raw data for all machine learning workflows with all concentrations converted to parts per million (ppm).

As part of the inferential data generation portion of the workflow, feature imputation was used on only features but not on data labels following Zhang et al. (2021a). Feature imputation was not applicable for the anomaly detection portion because there are no missing entries in the inferentially generated geochemical data. For model selection and tuning, we employed fourfold cross-validation using a grid search (Table 3). We employed the R^2 or the coefficient of determination (CoD) metric for model

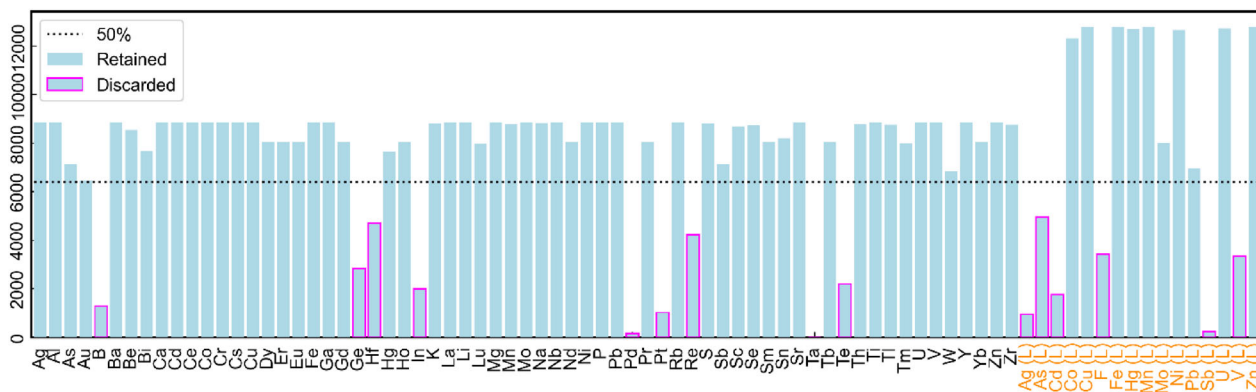


Figure 5. Absolute abundance (n) of elemental analyses in the training dataset. Elements that are in the legacy datasets are shown with a “(L)” following the element name and are in orange text color. The 50% population line (dotted line) is also shown and modern analyses that were below this line were removed from predictive modeling (purple bars).

tuning and the mean absolute percentage error (MAPE; e.g., de Myttenaere et al., 2016) to profile prediction performance. The mean absolute percentage error is the absolute value of the deviation in the predicted minus the actual quantity, divided by the actual quantity (similar to relative error). In Zhang et al. (2022b), the median absolute error (MedAE) metric was also used, which we forgo because all units are standardized into ppm for downstream usage, and therefore, there is a large quantitative range that does not facilitate comparisons between elements. A zone-based spatial cross-validation was used in Zhang et al. (2022b) to understand the performance implications when deploying models across zones. In this study, we also perform spatial cross-validation across 18 zones for which there is coverage of training data (all of Fig. 3b and c, and re-analysis sample locations shown in a). However, data coverage is variable with the least-sampled zone being NTS zone 064 K, which has a minimum of 5 samples for some elements (e.g., Ag). Hence, we also perform a weighted average to understand metric results in cross-validation. In our context, the focus is on application and model deployment into new zones. Hence, spatial cross-validation provides some constraints on the validity of the predicted results. Spatial domain (gridded and interpolated results) performance metrics are not used in this study, unlike that of Zhang et al. (2022b), because the maps produced in this study are intended for interpretation and as such, the interpolation parameters (of ordinary kriging) generally differ (e.g., variogram model parameters). Hence, a one-to-one comparison at the image level is not generally possible.

Non-machine Learning-based Geochemical Anomaly Detection

Principal component analysis (PCA) is one of several methods used as part of the current study to identify geochemical anomalies. First, inferential data are transformed by using the centered log ratio (CLR; Aitchison, 1982) for each element and sample. The PCA method was then applied to the transformed data to extract multivariate relationships and to aid in the discrimination between the insignificant/background and significant or anomalous (i.e., ore deposit) processes (Carranza, 2008; Zuo, 2011; Grunsky et al., 2014; Harris et al., 2015; Grunsky and de Caritat, 2020). Essentially, PCA re-coordinate data along axes of variability, and a selection of the most variable axes reduces the dimensionality of chemical coordinates. Hence, the regional geochemical (or lithological) variability can be captured by principal components, which can correspond to sample stoichiometries or equivalently, mineral compositions (Grunsky, 2010; Grunsky et al., 2014; Grunsky and de Caritat, 2020). Anomalousness of concentrations can then be proxied by regression residuals of an element against the dominant principal components. The delineation of dominant to minor principal components can be made in a number of ways, e.g., on a basis of eigenvalue or variability explained. In our application, because partial ground truth exists in the study area in the form of known elemental occurrences or deposits as documented by the Manitoba Geological Survey (Manitoba Mineral Resources, 2013), we used the number (two) of principal components that produced anomaly maps that best corroborated the

Table 3. Hyperparameters for the random forest algorithm in both machine learning workflows

Parameter	Range (data generation)	Range (anomaly detection)
Ensemble size	1000	1000
Maximum depth	11 to 15, unlimited	11 to 15, unlimited
Maximum features	3 to 8	3 to 9
Minimum samples per split	3 to 6, unlimited	3 to 6, unlimited
Minimum samples per leaf	2 to 5	2 to 5

regional ground truth in a qualitative manner. The delineation of foreground to background principal components (e.g., major lithological variability-background) was determined through analyzing the variability explained of each component, combined with a qualitative matching that maximized the contrast of known regional mineral occurrences in the resulting anomaly maps. Specifically, maps of regression residuals against the dominant principal components (subtraction of the lithological signal) were visually analyzed to determine the number of dominant principal components to maximize the match of hotspots with mineral occurrences or deposits in the area.

RESULTS

Inferential Data Generation

Inferential data generation performance was evaluated using a spatial cross-validation. During cross-validation, all elements were systematically explored across the 18 NTS zones except for Au, which exhibited < 0.1 CoD (R^2) during model selection, and was essentially unpredictable. Results of the CoD and MAPE metrics (Figs. 6 and 7) show that elements that were analyzed in both the legacy and modern datasets predicted the best. For elements that were only present in modern datasets, the prediction performance generally varies by zone. In some cases, zones that featured the lowest number of samples, on average, predicted the worst (e.g., NTS 064 K). Elements that exhibited a CoD metric value less than 0.2 include As, Be, Bi, Ca, Cd, Cs, Na, Nb, S, Sb, Se, Sn, Sr, W and Zr. However, elements that exhibited a MAPE value greater than 1 only include As, Th and Zr. A qualitatively identical and quantitatively similar behavior was observed in the proof-of-concept study (Zhang et al., 2022b) using a smaller dataset of only the Labrador and Saskatchewan samples (Fig. 3b and c). Several

known mechanisms of model error include: (1) the existence of anomalies (Fig. 8b), which creates an asymmetric distribution of prediction residuals (difference between prediction and actual quantities, e.g., Zhang et al., 2021a, 2022b); (2) limited data variability (Fig. 8a, b and c) due to quantization issues of low elemental concentrations near the lower detection limit (e.g., Pb, U and Th in Zhang et al., 2021b); (3) pervasive censoring (e.g., Ta, Pd, Au, Pt and B in Zhang et al., 2021a); and (4) training data generalizability/spatial variability (Fig. 8c; e.g., Cd, Cs, Cr, Nb and Na in Zhang et al., 2022b and also by sample lithology in Zhang et al., 2021b). In addition, there are suspected mechanisms of error that include: (1) the nugget effect (Fig. 8d); (2) elemental mobility; and (3) primary data accuracy and precision (Table 2; also see Zhang et al., 2021a, b, 2022b). For the relatively poorly predicted elements in this study, known anomalies in the training data (Zhang et al., 2021a) clearly affected the prediction performance of Be, Bi, Nb, Cs and Se. Limited data variability affected Be, Bi, W, Sn and Se. Training data generalizability affected Cd, Cs, Cr, Nb and Na. By design of the method (Zhang et al., 2022b), pervasively censored elements ($> 50\%$ records missing in this study, Fig. 5) were unpredicted. However, the amount of training data is different for all elements, with the most (left-) censored elements being Au, As, Sb and W (Fig. 5), which degrades their prediction performance. However, for our data, the effect of left-censoring would only be a significant contributor to model performance, if limited data variability due to data quantization is also a problem, because this combination results in a major loss of data variability.

The effect of primary data accuracy and precision (variable domain dimensions of scientific data quality) on machine learning is not yet studied. In this deployment study, we provide an analysis of the effects of primary data accuracy and precision on inferentially generated secondary geochemical data. For the accuracy and precision of primary geo-

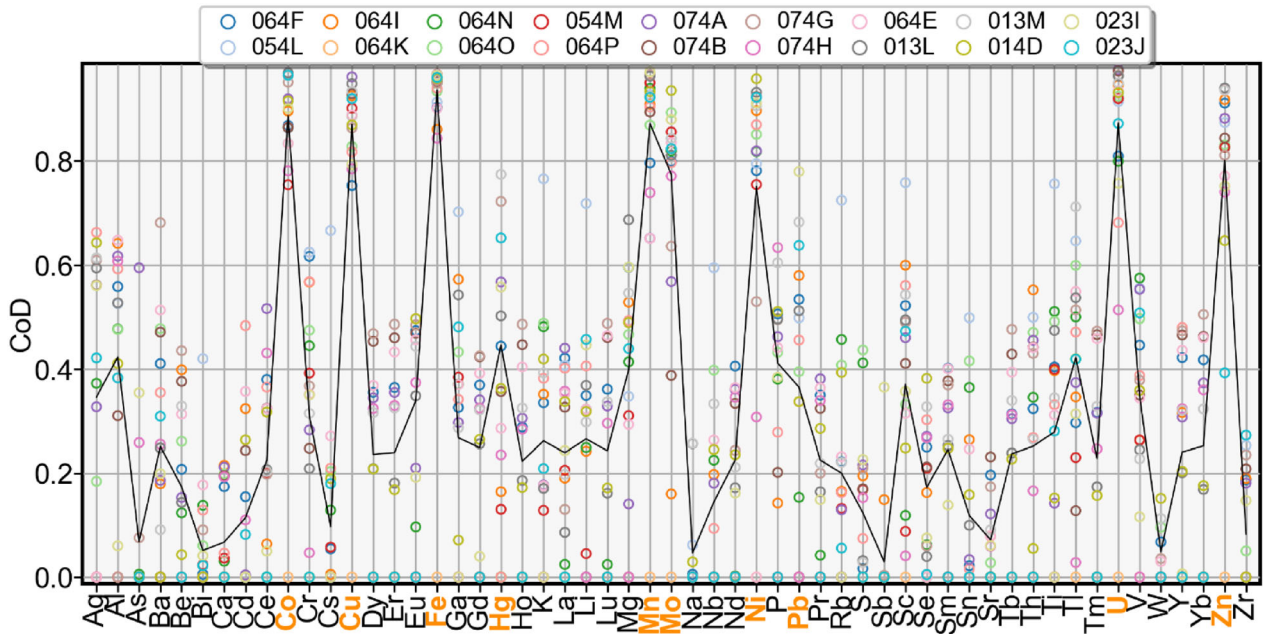


Figure 6. CoD (R^2) metric results of spatial cross-validation across 18 NTS zones in the training dataset (colored circles). The line shows the membership-weighted averages of each element. The orange-colored element labels are elements that are present in both legacy and modern datasets. Black-colored element labels are only present in modern datasets.

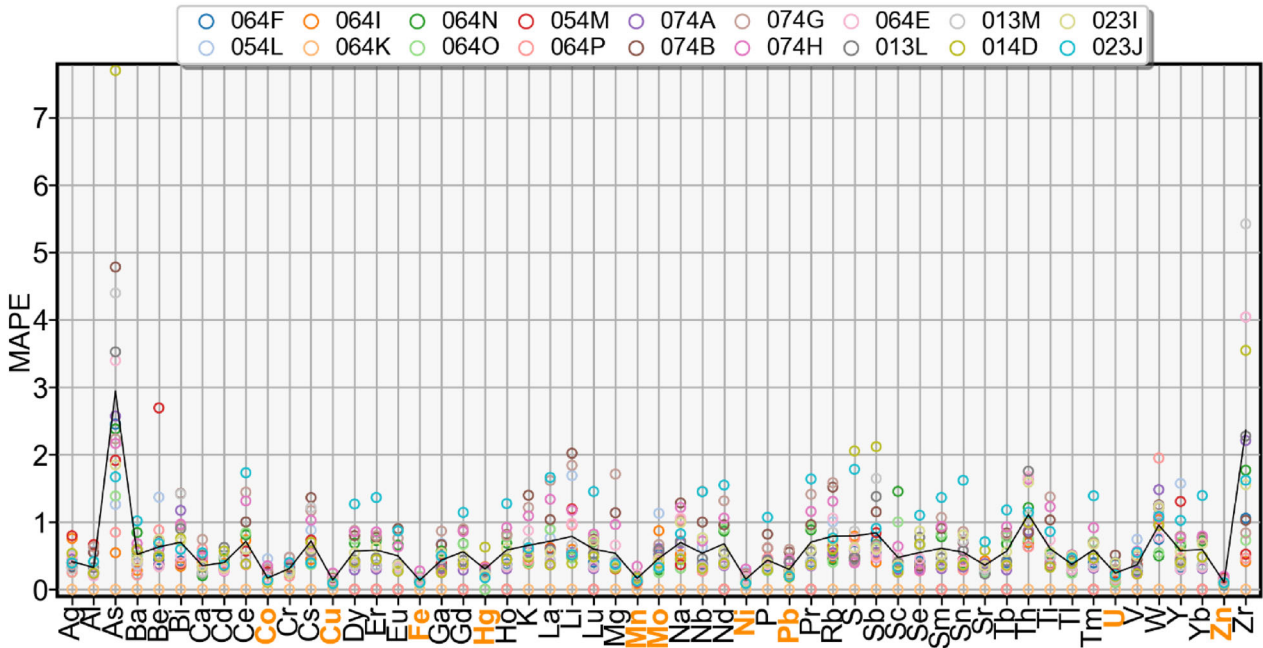


Figure 7. MAPE metric results of spatial cross-validation across 18 NTS zones in the training dataset (colored circles). The line shows the membership-weighted averages of each element. The orange-colored element labels are elements that are present in both legacy and modern datasets. Black-colored element labels are only present in modern datasets. Zone 064K did not feature sufficient samples to permit MAPE computation (shown as 0 values instead for visibility).

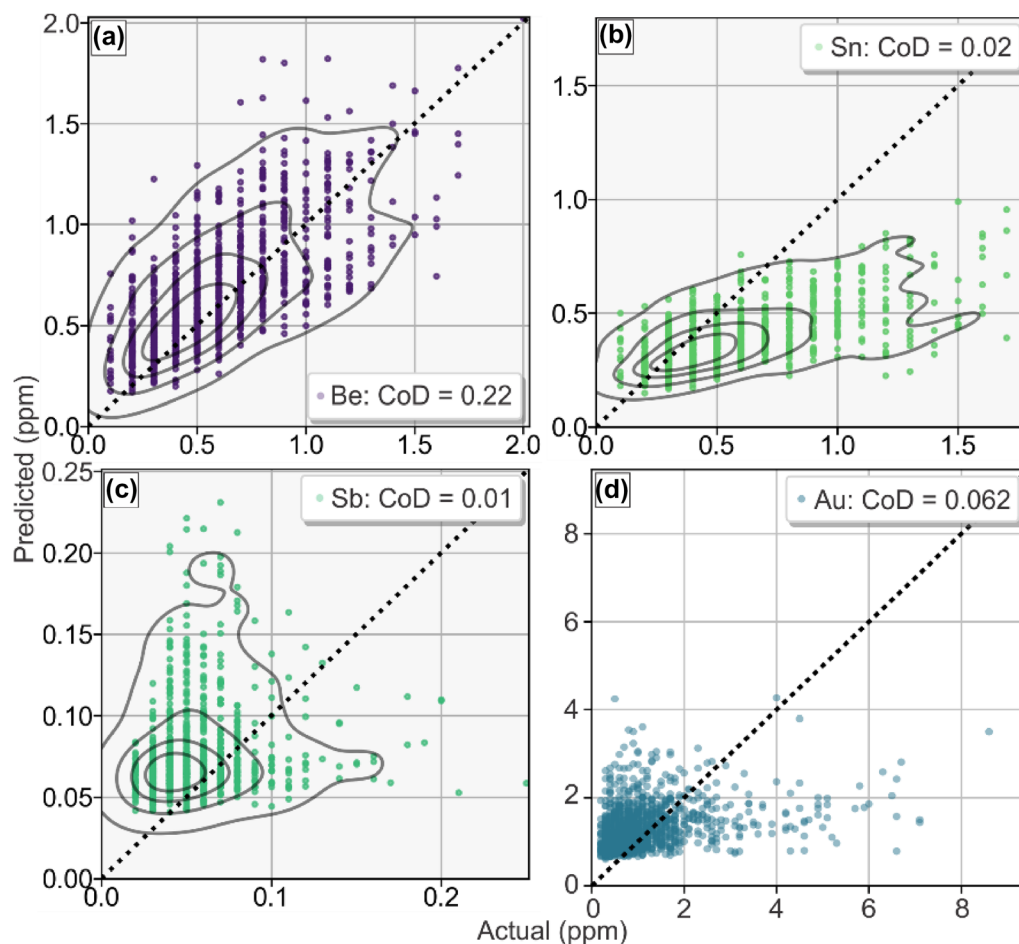


Figure 8. Notable issues observed during cross-validation (a–c) and during algorithm selection (d). (a) Limited data variability due to quantization caused by low elemental concentrations near the lower detection limit. (b) Quantization issues and existence of (physical or otherwise) anomalies that create an asymmetric spread of the scatter cloud. (c) Poor training data due to a combination of significant data quantization and high-dimensional variable domain issues (e.g., poor elemental relationships). (d) Potential significant nugget effect already noticeable during algorithm selection, leading to low model performance. CoD = coefficient of determination or R².

chemical data, we use the relative error (RE) and RSD metrics (McCurdy and Garrett, 2016). For accuracy of the predicted data, we use the average MAPE values for each element. The definition of the MAPE metric is the same as the RE metric if the analyzed quantity in the RE metric is replaced by the predicted quantity in the MAPE metric. The RE and RSD values of the primary geochemical data and MAPE for the inferentially generated, secondary geochemical data are given in Table 4. For most elements, the accuracy of primary and secondary geochemical data is within one order of magnitude (RE compared with MAPE). However, this is not the case for elements As, Mo and S, which

have substantially higher MAPE compared to RE and RSD. Furthermore, elements Be, W and Zr have RSD values greater than 20%, which could indicate laboratory accuracy issues or nugget effects (Table 4). It is important to understand the effect on secondary data accuracy that is induced by primary data accuracy and precision. For this purpose, we modeled the relationship between the MAPE and a quadrature of the RE and RSD (Fig. 9), which reveals a weak correlation. Therefore, accuracy of the inferentially generated data is at least weakly affected by the accuracy and precision of the primary data used in training. However, this relationship does not account for the quality of the machine

Table 4. Accuracy comparison between laboratory-derived (analytical) and predicted data for zone NTS 064F (fully re-analyzed and within deployment area) in northern Manitoba

Element	Analytical		Prediction	Element	Analytical		Prediction
	RE (%)	RSD (%)	MAPE (%)		RE (%)	RSD (%)	MAPE (%)
Ag	4.20	5.40	33.48	Mo	2.72	4.52	52.00
Al	9.60	3.64	22.43	Na	N/A	9.93	44.68
As	1.70	5.70	242.93	Nb	24.43	5.03	37.34
Au	62.8	42.2	N/A	Nd	6.41	6.61	39.15
Ba	7.50	4.50	21.97	Ni	7.70	4.90	15.59
Be	1.80	24.9	43.67	P	10.07	5.49	34.02
Bi	13.31	13.95	42.13	Pb	10.34	4.91	27.60
Ca	1.38	3.38	23.66	Pr	3.35	6.21	41.39
Cd	9.45	5.47	30.74	Rb	11.30	4.50	48.28
Ce	2.20	4.30	37.77	S	1.04	3.87	57.35
Co	17.0	4.60	14.27	Sb	10.05	5.72	72.66
Cr	12.6	4.60	30.50	Sc	3.50	8.20	34.14
Cs	8.45	5.26	44.73	Se	3.10	6.00	41.27
Cu	3.89	4.09	12.71	Sm	4.51	5.53	36.74
Dy	5.86	5.79	34.51	Sn	11.8	5.20	34.95
Er	3.78	5.50	35.78	Sr	5.10	4.70	24.89
Eu	1.03	6.58	37.10	Tb	5.49	8.16	39.69
Fe	8.03	3.37	11.36	Th	8.70	18.9	69.32
Ga	13.3	5.80	31.44	Ti	25.67	5.59	38.29
Gd	5.72	6.61	36.62	Tl	6.20	5.64	28.84
Hg	11.1	7.00	25.69	Tm	3.97	8.07	39.49
Ho	3.09	7.30	37.54	U	1.00	4.90	37.50
K	N/A	7.90	39.38	V	1.30	6.60	31.78
La	5.20	4.40	37.86	W	9.50	21.7	72.73
Li	17.1	5.40	54.70	Y	4.82	4.92	32.88
Lu	1.32	8.60	40.76	Yb	4.62	6.58	35.53
Mg	10.13	3.90	41.24	Zn	6.80	4.20	7.08
Mn	1.50	3.60	10.75	Zr	58.3	22.8	105.32

Laboratory-derived accuracy was measured using reference materials (LKSD-4) from the Geological Survey of Canada (McCurdy and Garrett, 2016)

RE Relative error, RSD relative standard deviation, MAPE mean absolute percentage error

learning features, which are legacy data and exhibits generally higher RSD and unknown RE (Table 2). The lack of RE information in legacy data prevented an effective analysis of this relationship. Regardless of the source of the error, the effect of the error in variable domain (e.g., of the predicted versus actual concentrations per-sample) is substantially attenuated if the data are used for mapping due to the support effect (averaging of an ensemble of points over an area). This was previously observed and quantified using a similar training dataset covering the Saskatchewan and Labrador areas (because these are the only regions with a large and contiguous spatial coverage, such that spatial comparisons could be effectively made, Zhang et al., 2022b). It was found that the relative change for the CoD metric is about 42.21% and for the MAPE metric is about – 69.24%, in the spatial domain in the form of

maps compared with that in the variable domain (Zhang et al., 2022b). This implies that concentration maps are more reliable than what is suggested by solely variable domain metrics, which do not take into account the usage of data.

Anomaly detection results

Since the vast majority of the deployment area in northern Manitoba is covered solely by legacy geochemical surveys (Fig. 3a), the geochemical anomalies in the area are enhanced/detected using two methods and are evaluated qualitatively using known locations of mineral occurrences. A data reconstruction-based anomaly detection using machine learning was performed on a total of 46 elements (mostly inferentially generated data, except

where modern data already exist, Fig. 3a), excluding Au, which could not be reliably predicted and 9 elements used as features during anomaly detection. Since the features were all major and minor elements, the 46 elements were all trace elements. Data reconstruction (model) performance was measured by the CoD and MAPE metrics (Fig. 10). Except for U and, to a lesser extent, As, Mo and W, the MAPE metric scores were excellent (< 0.2) (Fig. 10b). The CoD metric scores show a similar pattern, but the absolute differences between individual elements were smaller than that of the MAPE metric (Fig. 10a

compared with b). Higher scores generally imply that the anomalies would be more specific and spatially selective (Zhang et al., 2021a).

Using the CLR-transformed data (of a total of 55 elements, 46 elements overlapping with machine learning-based anomaly detection and 9 elements that were used as machine learning features), we systematically explored the regression residuals qualitatively by matching the observed anomalies with known mineral occurrences on a map to optimize the number of principal components used in the multilinear regression. Ground truth in the area in the form of known mineral occurrences or deposits was available for Co, Cu, Ni and U (Manitoba Mineral Resources, 2013). The explained variance of the principal components decays rapidly and visually reaches an elbow around principal component 6 (Fig. 11). This indicates that the large-scale variability in the data is best captured by principal components no larger than 6. For several other elements, and most notably Ag, W, Li, Bi and REEs, there are fewer known occurrences centered mostly in the southern portion of the study area, where the region is better explored (Davies et al., 1962). Systematically exploring multilinear regression residuals for Co, Cu, Ni and U showed that the optimal number of principal components is 2 (~54.13% of variability explained). Higher variability explained implies that the anomalies would be more specific and spatially selective. However, at more than 2 principal components, the anomalies that match with the ground truth begin to attenuate and some

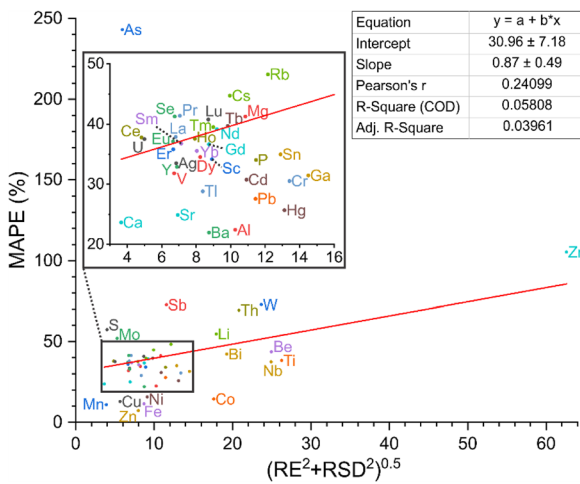


Figure 9. Mean absolute percentage error (MAPE) of predicted data versus the relative error (RE) and relative standard deviation (RSD) of laboratory-derived data and a fitted model. A weak correlation is observed.

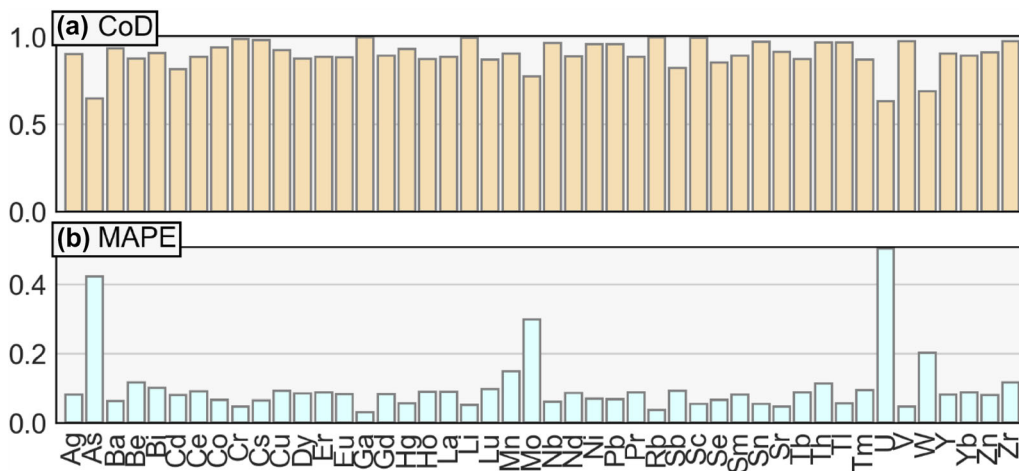


Figure 10. Metric scores for the CoD (R^2) and MAPE metrics for the random forest algorithm during anomaly detection.

become unrecognizable, possibly because these other principal components reflect more subtle geochemical patterns and processes. In contrast, the first two principal components likely reflect a strong lithological control that is important for predicting lake sediment geochemistry (discussed below).

It is possible to independently gain insight into the validity of the anomaly detection methods and remove the effects of inferential data generation, by examining solely the validity of the anomaly detection outputs using an elemental concentration that is

not inferentially generated. This allows us to isolate the effects of the inferential data generation from the effects of the anomaly detection. For this purpose, we examined the concentration and anomaly maps of Co, Cu, Ni and U (Ni shown in Fig. 12). Ni anomalies align reasonably well with known mineral occurrences/mines in the area (Fig. 12b and c; Manitoba Mineral Resources, 2013). The process and results were similar for the other three elements. These observations can be extended to elemental concentrations that were inferentially generated and whose ground truth is spatially similar in distribution (e.g., Ag, Fig. 13a-c). In this case, the spatial similarity is a result of multivariate geochemical anomalies that result from mineral occurrences hosting a variety of related minerals and/or mineral chemistries. For Ag, the predicted concentration and anomaly maps are also a general match with known Ag-bearing mineral occurrences or deposits. For the lesser-explored elements that are part of the Canadian critical minerals (Government of Canada, 2022), far fewer known occurrences and deposits exist for Bi, (Fig. 13d-f), Li (Fig. 13g-i), REEs (Fig. 14) and W (Fig. 15g-i). In all cases, predicted hotspots and anomalies generally match well with the known ground truth, given the sediment nature of the samples (that tend to mix through transport,

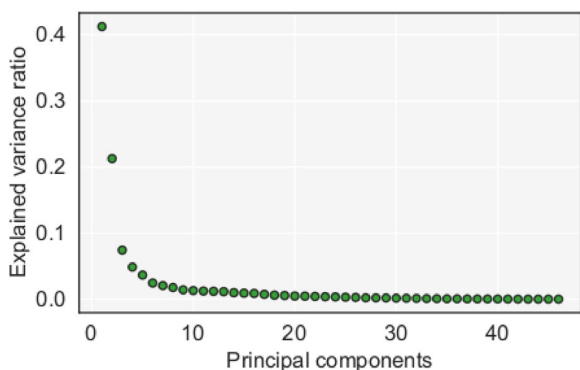


Figure 11. Scree plot of principal components of the CLR-transformed, primarily inferentially generated geochemical data.

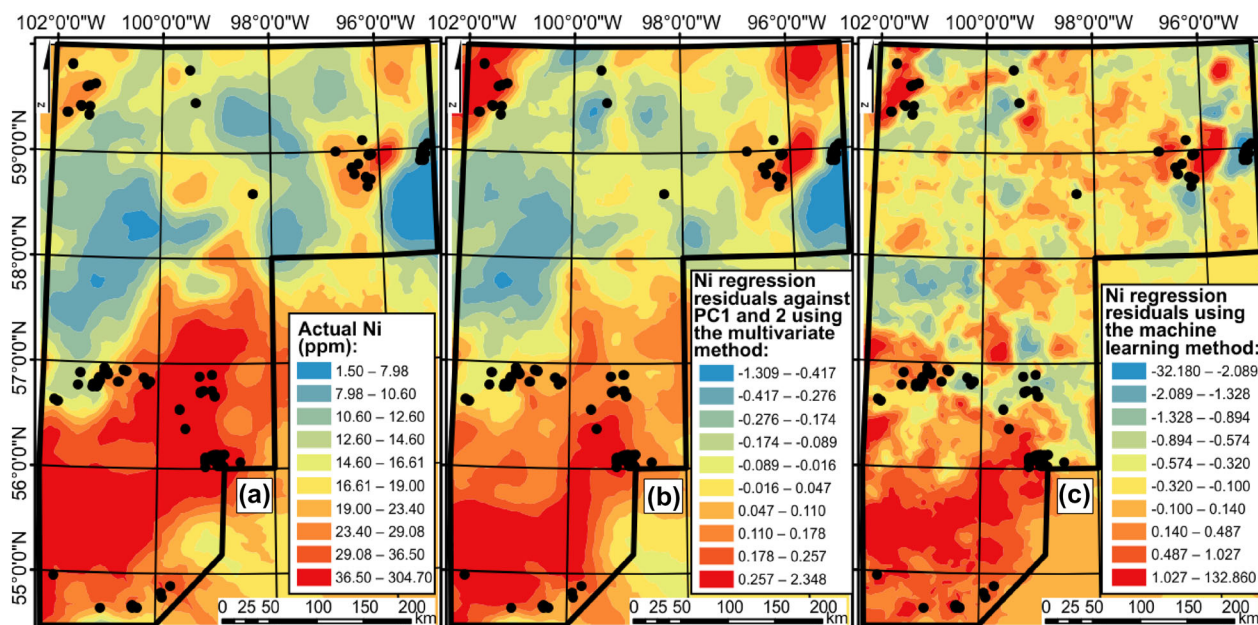
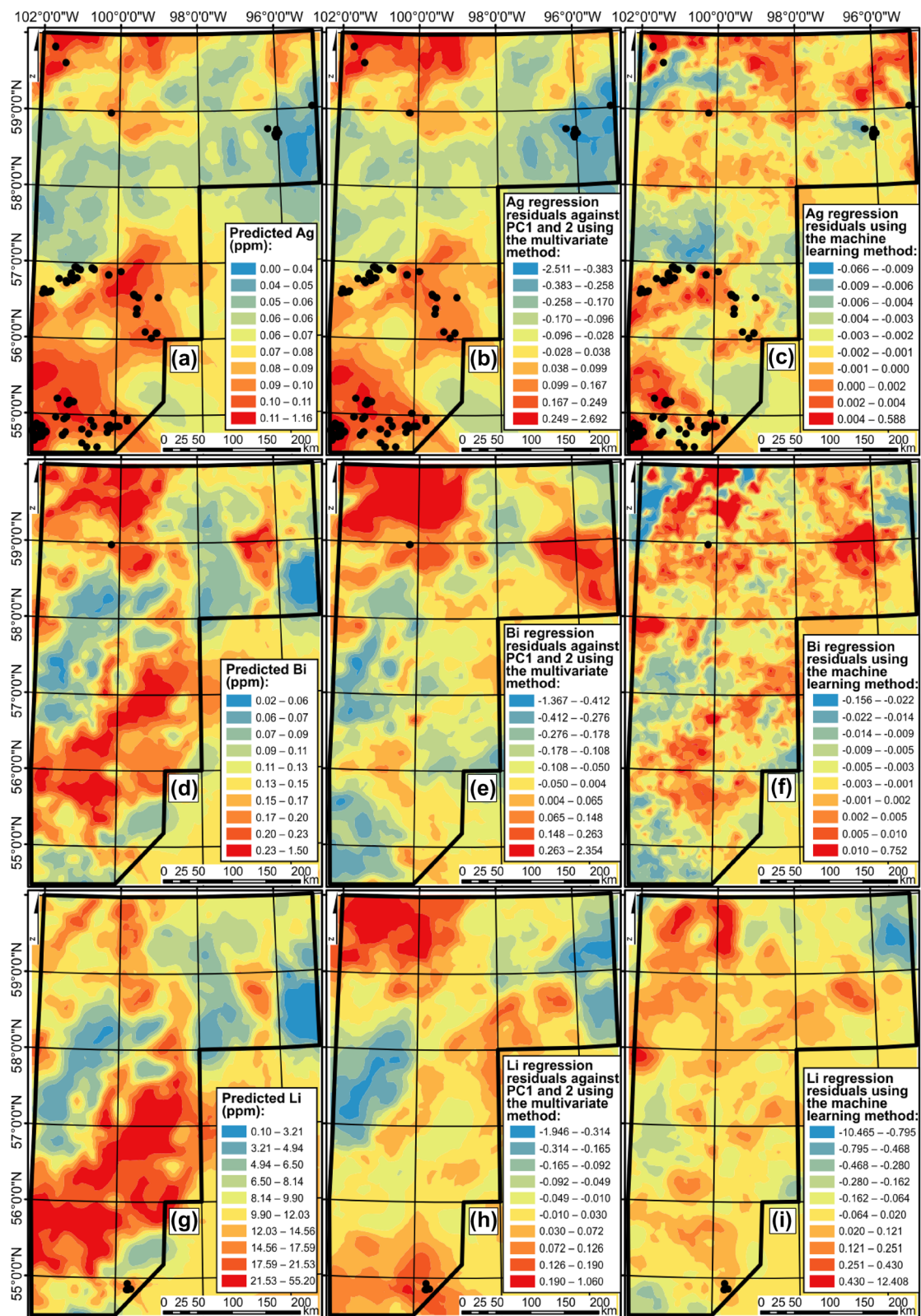


Figure 12. (a) Concentration maps of nickel (Ni; primary geochemical data, not inferentially generated) and anomalies as determined through two methods (b, c). The interpolation method used here is simple kriging. Hyperparameters for (a) are nugget = 0.44, sill = 0.74, range = 389.82 km, using the stable model (similar for other maps). The polygon (black line) marks the extent of sample coverage (also see Figs. 2 and 3a). The black circles correspond to known Ni occurrences or deposits in the Manitoba study area.



◀ **Figure 13.** Inferentially generated geochemical data for silver (Ag), bismuth (Bi) and lithium (Li), including their anomalies as determined through two methods (Ag [a to c], Bi [d to f] and Li [g to i]). The polygon (black line) marks the extent of sample coverage (also see Figs. 2 and 3a). Known occurrences or deposits for each element are shown as black circles.

which blurs trends). Hence, hotspots that do not align with known mineral occurrences can reasonably be expected to be explored further.

The mapping results of the inferentially generated geochemical data, PCA-based anomaly detection and machine learning-based anomaly detection, for select elements, are shown in Figures 13, 14 and 15. For each element, the anomalies in the study area as produced through the two anomaly detection methods have reasonable agreements. For example, in the REEs maps (Fig. 14), the northeast and mid-west portions of the map are consistently depicted as anomalous. These regions further overlap with elemental concentration hotspots. However, anomalies do not necessarily overlap with concentration hotspots. This is because both the machine learning-based and the principal components-based anomaly detection methods attenuate compositional variability of samples by removing a portion of the regional background concentration of elements (e.g., Zhang et al., 2021a, b). This effect is quite prominent in the Li maps, where a large anomaly exists at the northwestern corner of the study area as identified through both anomaly detection methods, but which is not an elemental hotspot (Fig. 13g, h and i).

INTERPRETATIONS

Our geochemical anomaly interpretation aims to create a list of proven to probable anomalies (*sensu lato*) and rank them by their level of uncertainty (detailed below; Fig. 16; Supplementary Information). This enables us to create a framework for predictive geochemical exploration, as opposed to knowledge-driven or systematic geochemical exploration. An outcome of predictive geochemical exploration is that exploration could be, in essence, data-driven, not excluding the use of knowledge. Not only is data-driven exploration more suitable for brownfield settings, but it is also more agile because sampling could become spatially targeted toward known anomalies. The objective of the framework is

to conduct geochemical reconnaissance using inferential data generation, followed by a goal to decrease exploration risks by spending resources to lower targeting risk. To accomplish this goal, we use the following risk categorization criteria:

Category 1. Low uncertainty or proven, where geochemical anomalies occur within the coverage area of most modern data (locations with samples in Fig. 3a).

Category 2. Moderate uncertainty or moderate probability, where geochemical anomalies are continuous into or out of the coverage of modern data or where its multivariate elemental associations (as determined through the PCA analysis) exist within modern data coverage.

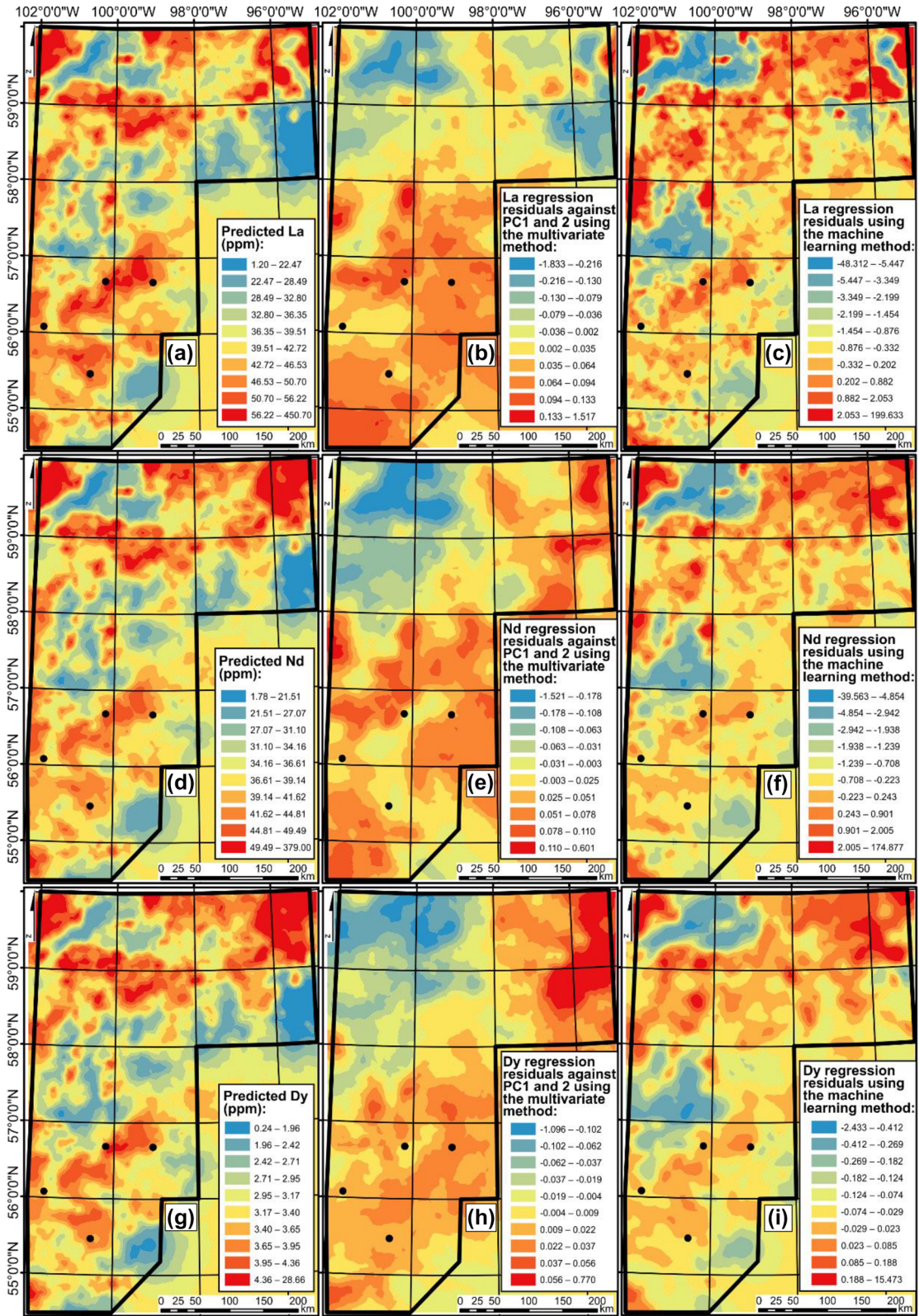
Category 3. High uncertainty or low probability, where geochemical anomalies occur within solely inferentially generated data.

We chose these categories to facilitate comprehension and exploration of all geochemical anomalies. In addition, this categorization scheme allows a future operationalization of non-grid-based and data-driven multi-resolution regional surveys. The operationalization of such surveys would be motivated by a priority to progressively move anomalies in Category 2 into Category 1 through more detailed local analysis, additional lines of evidence (e.g., remote sensing or geophysics), re-analysis of samples and, if necessary, additional sampling (or conversely, through the elimination of those anomalies where they do not yield occurrences). Similarly, a second (or lesser) priority would be to move anomalies in Category 3 into Category 2 (or alternatively directly into Category 1) or eliminate anomalies where they do not yield any occurrences. This essentially provides a modern alternative to a fixed resolution, grid-based reconnaissance, prospectivity and regional exploration.

DISCUSSION

Validity of the Results and Limitations

Algorithmic choice during inferential data generation affects the overall results. The random forest algorithm has been demonstrated to be suitable and, in some sense, desirable for the generative



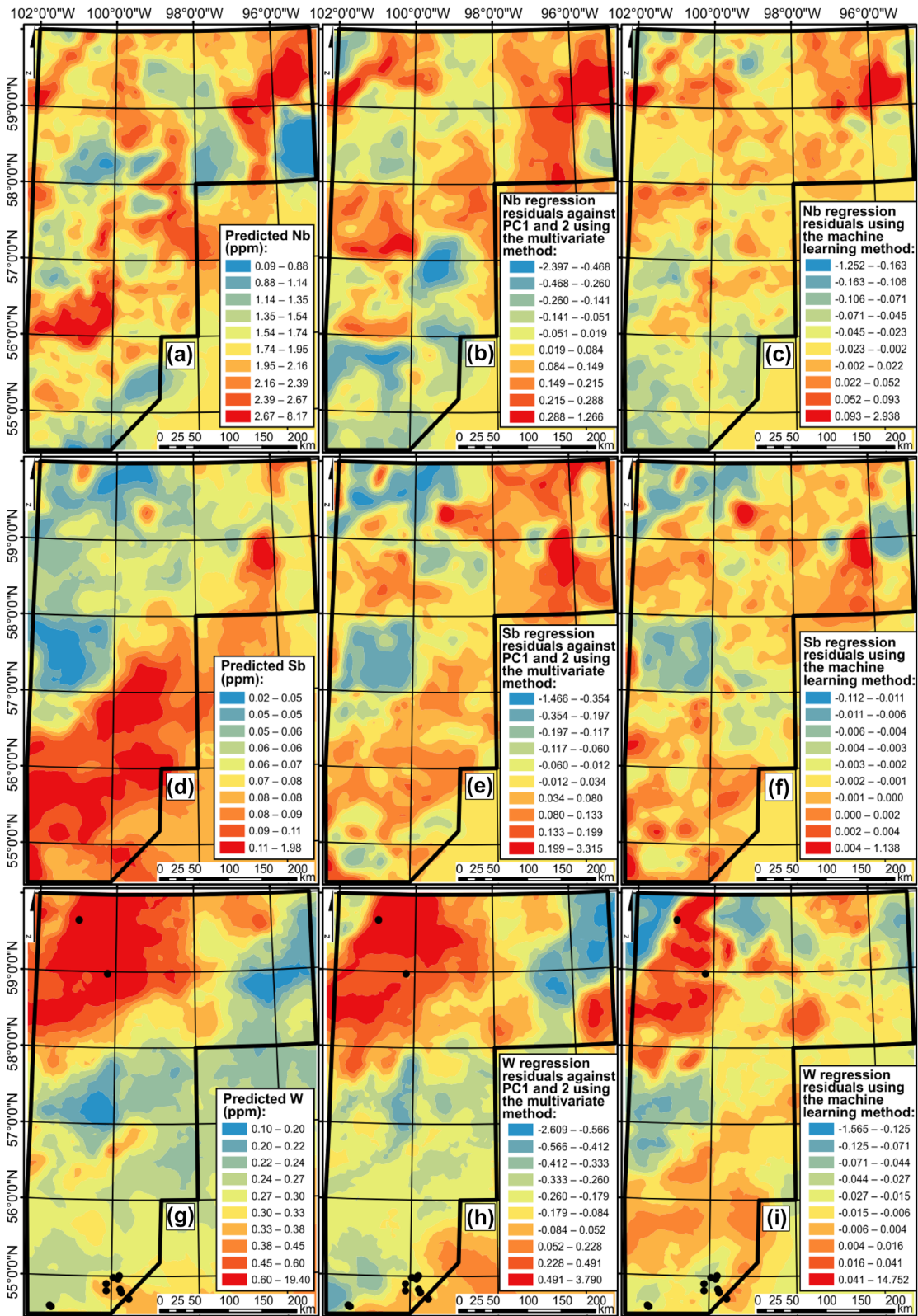
◀ **Figure 14.** Inferentially generated geochemical data for select REEs and their anomalies as determined through two methods (La [a to c], Nd [d to f] and Dy [g to i]). The polygon (black line) marks the extent of sample coverage (also see Figs. 2 and 3a). Known occurrences or deposits for each element are shown as black circles.

data task in geochemistry, among other shallow learning algorithms (Zhang et al., 2021a, b; 2022b). The desirability of this type of algorithm is traceable to its lack of geometry awareness in the feature space. This meant that data transformation or embedding (e.g., log-ratio transformations) is unnecessary, which was deduced theoretically and demonstrated empirically in Zhang et al. (2021a, b; 2022b). As a result of not necessitating embedding, a further implication is that imputation is generally not required, which is not the case for typical log-ratio transformations (Zhang et al., 2021b, 2022b). Consequently, the generated data are more reproducible across workflows. Lastly, because bagging overcomes the susceptibility of individual decision trees to noise, random forest, in comparison with other shallow learning algorithms, is relatively more resistant to overfitting, whereas boosting, for example, is much more intrinsically susceptible to noise (Li & Bradic, 2015). However, one of the important implications of using the random forest algorithm is that predicted concentrations must occur within the numerical range of the training data (both features and targets). For low-abundance elements and poor analytical methods, the algorithm has the potential to overestimate concentrations because all of the reported values are left censored. Measured concentrations for elements that are depleted relative to the continental crust and/or associated with rare or resistive minerals are the least likely to reflect their true values. In contrast, the predicted geochemical hotspots likely underestimate the true concentrations for true anomalies. As a result, it is possible that the maximum concentrations for each element are conservative in some cases. Despite these potential caveats, exhaustive spatial validation (Zhang et al., 2022b) suggests that random forest-based predictions are reasonable for most surveys and elements.

Prediction performance is highly dependent on the quality of the training data. The number of features used in data generation is potentially a limiting factor of the performance of inferential data generation. Due to standardization issues in legacy

data, the only consistently analyzed elements that met the criteria to be adopted as machine learning features amounted to 7. However, although the addition of more elements would mean that multivariate relationships would be better captured by the machine learning model, linearly increasing dimensionality of the feature space would require exponentially increased amounts of training data. This is known as the ‘curse of dimensionality’ and the positive effects of additional covariates would be ameliorated by the decrease in data density (e.g., Crespo Marquez, 2022). Hence, it is unpredictable whether additional features would increase model performance, at least without dimensionality reduction or feature selection techniques. Prediction performance is also dependent on the relationship of the training data with the target region, as discussed in detail in Zhang et al. (2022b). To maximize prediction performance, the training data should be as representative as possible of the target region, which will in turn increase model generalizability (e.g., class imbalance problem). This is particularly important given that the spatial or geostatistical (transferred) learning is not solved in the application of machine learning for geospatial data (e.g., Hoffmann et al., 2021). Nevertheless, there are two large categories of methods to analyze generalizability: (1) reductive approaches; and (2) data-driven approaches.

For reductive approaches, comparisons of the actual geological terrane between the training and deployment regions are necessary at some level of observation to ensure covariate comparability (Hoffmann et al., 2021). These levels could include: (1) field mapping; (2) sampling; and (3) microscopy and more reductive analytical techniques, such as micro-analytical and imaging techniques. Each refining scale of observation yields additional information but at substantially increasing cost and labor. At the field scale, all training data are contained within the Canadian Shield. An exact degree of similarity between sediments is not known, but the lake sediment surveys were originally designed to cover the Canadian Shield area and intended for integrated use (Friske, 1991; Bourdeau and Dyer, 2023). Furthermore, some of the training datasets are found directly within the target region (Fig. 3a). For these reasons, we consider the training dataset to be as representative at the regional scale as possible (especially given that this is the extent of data coverage currently available). At the sample scale, this analysis is generally impossible at the scale of



◀ **Figure 15.** Inferentially generated geochemical data for niobium (Nb), antimony (Sb) and tungsten (W), including their anomalies as determined through two methods (Nb [a to c], Sb [d to f] and W [g to i]). The polygon (black line) marks the extent of sample coverage (also see Figs. 2 and 3a). Known occurrences or deposits for each element are shown as black circles.

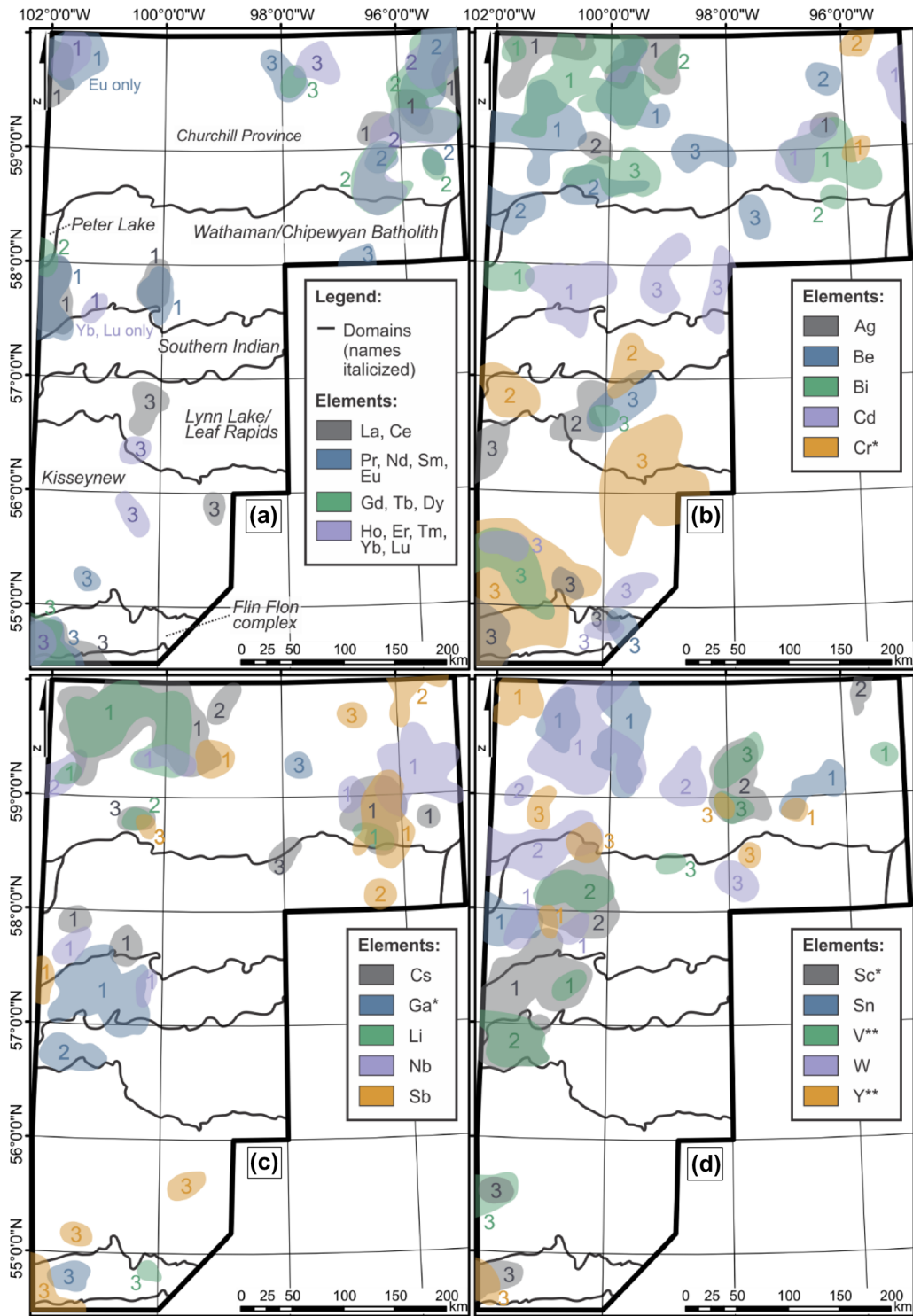
deployment, because lake sediment samples are very fine-grained and an in-depth analysis of sediment comparability is impossible to provide, given watershed variability in all areas. Similarly, in-depth analysis to ascertain bulk comparability between samples using even more reductive methods is impractical to achieve.

For data-driven approaches, there is machine learning-specific methodology that uses some form of block cross-validation to understand the extent of generalizability. This was fully analyzed for a mostly similar dataset in Zhang et al. (2022b), which used NTS zones as blocks to perform spatial cross-validation. A major finding was that for the purpose of mapping, which introduces the notion of support and therefore spatial averaging of samples, the performance was generally acceptable to very good in the spatial domain when comparing maps that were created using predicted versus actual data (Zhang et al., 2022b). In the variable domain, performance was not as consistent, which is also observed in this study (e.g., Figs. 6 and 7). However, because this study is not a method development study and is a deployment study, there is insufficient ground truth in the area; it is therefore not possible to further study the image domain performance by comparing maps of predicted versus actual data. We note that the methodology for spatial generalizability is presently incomplete because the metrics for spatial generalizability do not yet exist and this is an active area of research (e.g., Hoffmann et al., 2021). There are known issues (e.g., of covariate shift) associated with existing metrics and cross-validation strategies, such as block cross-validation. However, addressing these issues is beyond the scope of this study, but we foresee that our predictive data generation approach would be formally benefitted by the existence of a complete spatial learning methodology.

The possibility that element concentrations are either underestimated or overestimated raises two contrasting philosophies that result in polar opposite outcomes—over- or under-prediction as a default choice. The choice affects the balance between risk and reward downstream. The deployment context and the purpose of the workflow essentially wholly

control the suitability of either philosophy. A conservative approach would be to adopt under-prediction at the risk of a reduced number of geochemical hotspots or anomalies. This results in a reduced risk of downstream usage of such insights but potentially reduces the discovery probability of mineral deposits. The opposite outcome promotes a greater discovery probability of mineral deposits but increases the number of false positives anomalies. The adoption of either philosophy directly controls the suitability of various machine learning algorithms. For our purpose, as this is the first deployment of the ideas of Zhang et al. (2021a, 2022b), we erred on the side of caution, adopting a conservative approach. Hence, the inability of the machine learning algorithm to extrapolate outside of the training data is not just inconsequential, but rather desirable.

The limitations associated with both the inferential data generation and the machine learning-based anomaly detection methods were discussed fully in Zhang et al. (2021a, 2022b). As this paper does not focus on method development, we instead focus on the limitations associated with our proposed integration method and interpretation. An obvious limitation associated with our proposed integration of machine learning into data generation is that it relies on the existence of legacy data. Large-scale and decadal survey programs routinely conduct re-analyses and re-sampling to modernize legacy data, and hence, legacy data are an inevitable outcome. However, it is difficult to imagine that a sufficient amount would be generally available at smaller survey scales. Hence, as with all other data-driven methods, if training data are insufficient, of poor quality or no longer relevant, inferential data generation leveraging legacy data as an intermediary may be impractical or impossible. In this case, a remote sensing-based reconnaissance approach (including inversion of remote sensing data to geochemical data, e.g., Zhang et al., 2023) may be better, ground cover permitting. Another limitation of the uncertainty categorization is that moving targets from a higher uncertainty category to a lower one will not necessarily result in the discovery of a deposit. This process is only intended to guide exploration operations by creating clear data-driven survey objectives in the context of non-sequential, non-grid-based surveys. In any case, geochemical knowledge has only ever probabilistically led to mineral deposit discoveries, which is never guaranteed.



◀**Figure 16.** Location of the most prominent geochemical anomalies for predicted elements within the Manitoba study area. Numbers associated with each anomaly correspond to the uncertainty (risk) category. The geochemical anomalies were constructed by combining the PCA and machine learning anomaly detection methods. Overlapping areas with regression residuals falling within the two highest quantiles were drawn, up to a maximum of the ten-most prominent areas. REEs were qualitatively grouped based on anomaly map similarity. Note: *Only contains anomalies using the principal components-based multilinear regression model; **Anomaly locations did not match well when combining both anomaly detection methods.

Geological Interpretations

The benefits of compartmentalizing uncertainty at the data generation stage using inferential data generation and subsequently using a categorization scheme to rank anomalies imply that geological interpretation can be assessed in terms of uncertainty as well. Geological interpretations that are associated with Category 1, for example, are intrinsically more worthy of consideration than those in Categories 2 and 3. The ability to tie qualitative geological knowledge (e.g., geological maps) with a definitive uncertainty categorization is a major benefit of our approach. Here, we present our interpretations of a few select elements (Ni, REEs, Li and W), for which there is at least some ground truth in the form of mineral occurrences or deposits, exploring the relationship between anomalies, known mineral occurrences/deposits and the regional geology. Numerous other interpretation possibilities also exist, such as grouping certain elements to explore for known ore deposit types. For example, elements Li, Rb, Cs, Be, Ga, Sc, Y and REEs can be combined together to explore for REE-granitic pegmatites (Černý, 1991). This study is not meant as an exhaustive exploration of all possible geological interpretations and readers are encouraged to use the method/results from this study to derive their own geological interpretations.

Nickel

In 2020, Manitoba produced 8.6% of Canada's nickel, mostly from the Thompson Nickel Belt, which is located at the western-most edge of the Superior Province (Fig. 1; Manitoba Natural Resources and Northern Development, 2021a). Addi-

tionally, a significant number of occurrences/deposits have been documented south of latitude 57 (Fig. 12). Nickel occurrences/deposits further north are fewer. This is true of all elements in terms of exploration coverage, as the northern Manitoba region is less physically accessible due to season and terrain factors as compared with the southern Manitoba region (Davies et al., 1962). In our case, because Ni is not inferentially generated and the ground truth is better documented than for many other elements, the reliability of Ni anomalies provides a level of confidence in the anomaly detection methods. This is particularly desirable for the northern Manitoba region, because of sparse ground truth there. Both the PCA and machine learning methods have highlighted common geochemical anomalies in the study area, which spatially correlate very well with known occurrences/deposits as expected. No consistent surficial dispersal mechanisms are obvious from the geochemical anomaly maps (Fig. 12b and c), suggesting that local effects, such as drainage, could be the dominant transport mechanism of Ni-bearing sediments from known occurrences/deposits. In the interest of exploring for new deposits, three significant anomalies (shared between the two anomaly detection methods) are present in the Churchill Province (Fig. 12b and c). These anomalies were generated with actual data (non-predicted data), and therefore all have low uncertainty levels (Category 1). Two of the detected anomalies are associated with known occurrences/deposits which may also include Cu, Au, Pb, Ag and Zn (and Co for the westernmost anomaly). A few documented Ni occurrences, found immediately northwest from the dolomitic limestone formations, which rim Hudson Bay, have not been detected as anomalies (actual lows). High amounts of carbonate material can create potential issues due to high lake alkalinities, which can, in turn, affect the distribution and occurrence of elements in lake sediments (Karrow and Geddes, 1987). Interestingly, the northeastern-most detected anomaly is not associated with known occurrences and could be promising for new nickel prospects.

Rare Earth Elements (REEs)

In Manitoba, REE deposits have been found south of latitude 57, in the Trans-Hudson Orogen, and are associated with pegmatites, carbonatites and alkaline intrusions, containing between 0.36 and 16.8

wt.% REO + Y₂O₃ (Fig. 14; Manitoba Natural Resources and Northern Development, 2021b). In our investigations, several anomalies have been noted in the study area. Notably, known occurrences match relatively well with anomalies detected using both the multivariate and machine learning methods (Figs. 14 and 16a). Interestingly, the anomalies associated with known occurrences occur in areas where the data were predicted and are thus associated with a high risk (Category 3). But most importantly, the anomalies associated with known occurrences are not as strong as compared to those recorded further north, perhaps due to their nature. Pegmatites, carbonatites and alkaline intrusions tend to occur as small igneous intrusions. Weathering of these small intrusions may produce subtle geochemical signatures depending on whether the host rocks are differentiated or otherwise REE-rich. The strongest anomalies are found further north, overlapping with the Reindeer Lake (Fig. 2) or at the far northeastern and northwestern edges of the study area in the Churchill Province (Fig. 16a). Although there is sparse knowledge in the area, the majority of the detected anomalies are associated with a low-medium uncertainty (Category 1 and 2). The Churchill Province is known to host REE (+U, Th) deposits (e.g., Kulyk Lake REE occurrence; Saskatchewan Geological Survey, 2018). Furthermore, the predominant REE anomalies overlap between several elements (e.g., the north eastern edge of the study area contains anomalies of all REE elements), yielding additional credibility to our interpretations. Therefore, these lines of evidence are quite promising for the discovery of new REE prospects in northern Manitoba.

Lithium (Li)

In Manitoba, several lithium deposits have been found south of latitude 55 (Davies et al., 1962; Manitoba Natural Resources and Northern Development, 2021c), notably in the Superior Province and the eastern portion of the Flin Flon domain (as seen in Fig. 13g, h and i). In our investigations, several anomalies have been noted in the study area (Fig. 13h, i and 16c). The anomaly noted in the eastern portion of the Flin Flon domain matches quite well with known occurrences/deposits in the area, which includes the Green Bay (1Mt at 1% LiO₂; Powell, 2019) and Snow Lake prospects (11.1Mt at 1% LiO₂; Snow Lake Lithium, 2022).

Both prospects correspond to Li-bearing pegmatites. However, the most notable and strongest anomalies are found in the northwest Churchill Province (Fig. 16c). Interestingly, although there is sparse exploration in the area, the detected anomalies are associated with low uncertainties (Category 1). Furthermore, both the Superior and Churchill Provinces share similar geological contexts (Hanmer et al., 2004), with a number of notable deposits having been discovered and exploited in the Superior Province extent in Manitoba (Manitoba Natural Resources and Northern Development, 2021c). Lastly, the anomaly denoted in the eastern portion of the Flin Flon domain matches well qualitatively with known occurrences/deposits, therefore yielding additional credibility to the anomalies detected in the northwest Churchill Province. Thus, the strongest anomaly detected in the northwest Churchill Province is very promising for new lithium prospects.

Tungsten (W)

In the study area, most detected anomalies are found in the mid to northern regions (Fig. 16d). The largest and strongest anomaly is located in the northwestern portion of the Churchill Province. Two known occurrences overlap quite well with the largest and strongest anomaly detected, adding credibility to our interpretations. Interestingly, most of the detected anomalies are associated with low to moderate uncertainties (Category 1 and 2). These lines of evidence, when combined with the fact that there is sparse exploration in the north, are promising for future tungsten exploration as veins, skarns and/or possibly intrusion-hosted mineral systems.

Implications for Geochemical Exploration

A new approach to geochemical survey design is urgently needed as shallow and easily discovered mineral deposits are becoming harder to find and geochemical survey is at risk of becoming superseded by other techniques. Geochemical exploration can still evolve and modernize by leveraging its key competitive advantages: (1) that it provides chemical/elemental information, which is a stronger and more direct proxy of mineral deposits than remote sensing or geophysical data, both of which require

inverse modeling for further interpretation (Tarrantola, 1984; Srivastava et al., 2020); (2) geochemistry is less ambiguous than spectral data, meaning that it can be interpreted with higher confidence; and (3) due to continual improvements in analytical instrumentation (e.g., Table 2), large survey programs and companies likely possess copious amounts of legacy data, which readily suits data-driven predictive exploration through inferential data generation as demonstrated in this study.

In this study, we proposed an alternate brown-field geochemical exploration framework that leverages data-driven predictive targeting, essentially creating a new domain that we term 'predictive geochemical exploration.' In this new framework, integration of predictive analytics (including artificial intelligence and machine learning) plays an essential role to provide a shortcut to primary data generation by providing a nearly zero-cost reconnaissance of probable inferred (secondary) data. Consequently, targeting decisions can be made in a very rapid manner because timescales associated with inferential data generation is vastly shorter than primary data generation. A key benefit of this style of agile and multi-scale predictive exploration is that it de-risks further physical exploration by providing probable exploration targets. Therefore, subsequent local surveys are more targeted, data-driven, cheaper and timelier than regional-scale grid-based surveys.

Adoption of predictive geochemical exploration into existing survey programs is conceptually simple, and there are many possible approaches. The least invasive approach would be only to use predictive targeting to sequence grid-based surveys, such that the resulting survey data are still fixed resolution without a sampling bias toward probable hotspots or anomalies. This approach requires no major changes in existing survey programs and only deployment of trained models, and legacy data are necessary. In other words, there is a significant potential for a stronger integration of data-driven survey sequencing. Where legacy data do not exist, it would be necessary to produce a large training database through some combination of existing samples, or preferably, conduct a new sampling campaign. Alternatively, the most transformative approach would be only to conduct multi-scale and data-driven surveys, using multiple iterative passes. In this manner, hotspots and anomalies are preferentially sampled at a higher spatial resolution than the general regional background.

Implications for Prospectivity Mapping and Data-driven uses of Survey Data

Discontinuous data coverage is one of the fundamental problems in geological prospectivity modeling, rendering the models biased (Zuo et al. 2015; Council for Geoscience, 2022). This problem cannot be solved using interpolation or traditional data imputation techniques, because spatially missing data generally may not have covariates. Interpolation techniques also are not generally useful because missing data may span large spatial regions that are beyond, for example, the effective range of variograms that are used for kriging. The use of other types of interpolation, such as parametric techniques, sacrifices physical realism for simplicity, and the interpolated or extrapolated data can range from offering little to no scientific value to completely misleading. Instead, the methodology proposed by Zhang et al. (2022b) and applied herein fill in some data gaps, where covariates exist in the form of legacy data, producing predictions that comply with the inherent data characteristics and complexity of geochemistry, thereby providing a possible solution to the problem of bias.

Anomaly maps generated with inferential geochemical data are spatially correlated with the location of mineral occurrences (Figs. 13, 14 and 15), demonstrating the reliability of the applied methodology. This increases the confidence of using inferentially generated data in prospectivity mapping workflows. Although the study is a demonstration of method deployment (as opposed to development), the partial brownfield setting in the deployment area provides additional feedback on the validity of the method. This is especially beneficial because mineral prospectivity mapping uses a combination of data layers that could include geochemical data. A high degree of internal consistency between various data layers (e.g., between geochemical and mineral occurrence data) translates into derived prospectivity maps that are more reliable, since spatial consensus between data layers is more probable. Future studies, therefore, should focus on generating inferential data in additional deployment studies and advancing overall workflow performance (of the primary data used for training, the algorithms and models, and all other components) for the purpose of inferential data generation. Lastly, inferential data generation need not be restricted to geochemistry. Multiple other domains of data could be inferentially generated, provided that

reliable training datasets are available. For instance, for any type of survey program that cyclically modernizes its data collection through the deployment of newer instrumentation or sensors, it would be variably feasible to explore inferential data generation.

CONCLUSIONS

Geochemical surveys are one of the foundational datasets responsible for a number of mineral deposit discoveries and for geometallurgical uses (e.g., through element-to-mineral conversion). Advances in transdisciplinary scientific methods are creating opportunities to re-think survey design and assess whether machine learning can be deployed at various data pipeline stages. The integration of machine learning in geoscientific activities tended to (almost entirely) occur in the analysis and use of traditional geoscientific data. This is a key weakness in the innovation aspect of the use of transdisciplinary methods in geosciences because such focus on data analysis completely ignores the possibility to integrate machine learning into data generation. Hence, we address this major gap by demonstrating a dual-use integration of machine learning in a typical geochemical exploration pipeline to both predict modern multi-elemental geochemical data from low-dimensional and obsolete legacy analyses and to detect geochemical anomalies. To the best of our knowledge, this is the first study that demonstrates these novel aspects in a method deployment (not method development) context.

Here, we provided the first high-resolution probabilistic geochemical maps of the northern Manitoba region, before actual analyses or re-surveys have even occurred. The results demonstrate how inferential data generation can be used to predict the concentration of elements that were not included in the original survey design, including for a suite of critical minerals, for which, traditional exploration would take substantially longer time. Cross-validation metrics (MAPE and CoD) across 18 NTS zones in the training data suggest good performance for some elements (e.g., Ag, Li and REEs) and worse poor performance for others (e.g., Au, As, Bi, Cs, Sb, W and Zr). This underlying causes of this variable performance are multi-faceted scientifically and may reflect the importance of rare and/or resistive host mineral phases, left-censored data and/or other analytical issues with the underlying data. The good agreement between the

predicted lake sediment compositions and mineral occurrences is used as an extra validation measure for areas entirely devoid of modern geochemical surveys. Measured and/or predicted element concentrations, coupled with the PCA and the machine learning methods, were then used to identify and characterize geochemical anomalies with three categories of uncertainty. The predictive certainty-based exploration framework is a second novel contribution of this study. Our approach is purposefully not grid-based, fully leverages and rejuvenates existing data assets (including mainly obsolete legacy data) within survey programs and provides an agile and rapid method to discover mineral deposits. The proposed framework is also a natural extension of grid-based exploration programs because it realizes the data value that had been provided by systematic traditional surveys. In this framework, geochemical anomalies with low uncertainty are based, in part, on modern geochemical survey results, including remote parts of northwestern Manitoba that are favorable for REEs, Li and W. In contrast, REE geochemical anomalies in southwest and northeast Manitoba are associated with higher uncertainty categories because of the reliance on predicted data. The exploration activity is thereafter directed based on the incrementation of certainty of valuable but uncertain targets into higher levels of certainty.

Combining machine learning at multiple stages of the data pipeline can thus be used to focus new geochemical surveys to areas that reduce the level of uncertainty for the most important element anomalies. This new data-driven and targeted approach enabled by the dual-use of machine learning and an uncertainty-based exploration framework has the potential to significantly reduce the cost and time required for geochemical exploration and provide an alternative future of higher operational agility and exploration speed for existing survey programs. Our results also provide another example of how legacy data can be of tremendous value and, in this case, fully re-purposed to address modern mineral exploration priorities in a timely and cost-effective manner.

ACKNOWLEDGMENTS

The authors would like to thank Richard D. Dyer (Lake Sediment Geoservices) for providing

crucial information regarding the particularities of lake sediment samples. We thank Patrice de Caritat (Geoscience Australia) and two other anonymous reviewers for their insightful comments that have greatly improved this study. The authors would like to thank E.J.M. Carranza for editorial handling. We also thank Stephen J.A. Day (Geological Survey of Canada) for help with the preparation of re-analysis data used in this study. Finally, we thank Danny Wright (Geological Survey of Canada) for his support.

FUNDING

Open Access funding provided by Natural Resources Canada. This study was supported by a Department of Science and Innovation (DSI), National Research Foundation (NRF), Thuthuka, Grant (Grant UID: 121973).

DECLARATIONS

Conflict of Interest The authors declare that they have no known competing financial interests of personal relationships that could have appeared to influence the work reported in this paper.

OPEN ACCESS

This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

SUPPLEMENTARY INFORMATION

The online version contains supplementary material available at <https://doi.org/10.1007/s11053-023-10273-6>.

REFERENCES

- Aitchison, J. (1982). The statistical analysis of compositional data. *Journal of the Royal Statistical Society. Series B (Methodological)*, 44(2), 139–177.
- Alf rez, G. H., Esteban, O. A., Clausen, B. L., & Ardila, A. M. M. (2022). Automated machine learning pipeline for geochemical analysis. *Earth Science Informatics*, 15(3), 1683–1698.
- Bergen, K. J., Johnson, P. A., de Hoop, M. V., & Beroza, G. C. (2019). Machine learning for data-driven discovery in solid Earth geoscience. *Science*, 363(6433), eaau0323.
- Booyens, R., Gloaguen, R., Lorenz, S., Zimmermann, R., Andreani, L., & Nex, P. A. (2019). The potential of multi-sensor remote sensing mineral exploration: examples from Southern Africa. In *IGARSS 2019-2019 IEEE International Geoscience and Remote Sensing Symposium* (pp. 6027–6030). IEEE. <https://doi.org/10.1109/IGARSS.2019.8899081>.
- Bourdeau, J. E., & Dyer, R. D. (2023). Regional-scale lake-sediment sampling and analytical protocols with examples from the Geological Survey of Canada. *Geological Survey of Canada, Open File*, 8980, 41. <https://doi.org/10.4095/331911>.
- Breiman, L. (1996a). Bagging predictors. *Machine Learning*, 24(2), 123–140.
- Breiman, L. (1996b). Stacked regressions. *Machine Learning*, 24(1), 49–64.
- Bureau Veritas. (2022). 2022 Integrated Report: For responsible progress. Retrieved April 14, 2023, from <https://group.bureauveritas.com/newsroom/2022-integrated-report-responsible-progress>.
- Cameron, E. M. (1994). Lake sediment sampling in mineral exploration. In M. Hale & J. A. Plant (Eds.), *Drainage Geochemistry in Mineral Exploration* (Vol. 6, pp. 227–267). Elsevier Science. <https://doi.org/10.1016/B978-0-444-81854-6.50013-4>.
- Carranza, E. J. M. (2008). *Geochemical anomaly and mineral prospectivity mapping in GIS*. Elsevier.
-  erny, P. (1991). Rare-element granitic pegmatites. Part I: Anatomy and internal evolution of pegmatite deposits. *Geoscience Canada*, 18(2), 49–67.
- Corrigan, D., Galley, A. G., & Pehrsson, S. (2007). Tectonic evolution and metallogeny of the southwestern Trans-Hudson Orogen. In W. D. Goodfellow (Ed.), *Mineral deposits of Canada. A synthesis of major deposit-types, District Metallogeny, the evolution of geological provinces, and exploration methods* (Vol. 5, pp. 881–902). Geological Association of Canada Mineral Deposits Division.
- Corrigan, D., Hajnal, Z., N emeth, B., & Lucas, S. B. (2005). Tectonic framework of a Paleoproterozoic arc-continent to continent-continent collisional zone, Trans-Hudson Orogen, from geological and seismic reflection studies. *Canadian Journal of Earth Sciences*, 42, 421–434.
- Corrigan, D., Pehrsson, S., Wodicka, N., & de Kemp, E. (2009). The Paleoproterozoic Trans-Hudson Orogen. A prototype of modern accretionary processes. In J. B. Murphy, J. D. Keppie, & A. J. Hynes (Eds.), *Ancient Orogens and modern analogues* (Vol. 327, pp. 457–479). Geological Society of London.

- Council for Geosciences. (2022). *Annual report, 2021/22* (p. 162). Republic of South Africa.
- Cover, T., & Hart, P. (1967). Nearest neighbor pattern classification. *IEEE Transactions on Information Theory*, *13*, 21–27. <https://doi.org/10.1109/TIT.1967.1053964>.
- Crespo Márquez, A. (2022). The Curse of Dimensionality. *Digital Maintenance Management*. Springer Series in Reliability Engineering. Springer.
- Davies, J. F., Bannatyne, B. B., Barry, G. S., & McCabe, H. R. (1962). *Geology and mineral resources of Manitoba* (p. 221). Department of Mines and Natural Resources, Mines Branch.
- Daviran, M., Parsa, M., Maghsoudi, A., & Ghezlbash, R. (2022). Quantifying uncertainties linked to the diversity of mathematical frameworks in knowledge-driven mineral prospectivity mapping. *Natural Resources Research*, *31*(5), 2271–2287.
- Davis, W. J., Hanmer, S., & Sandeman, H. A. (2004). Temporal evolution of the Neoproterozoic Central Hearne supracrustal belt: rapid generation of juvenile crust in a supra-subduction zone setting. *Precambrian Research*, *134*(1–2), 85–112.
- de Myttenaere, A., Golden, B., Le Grand, B., & Rossi, F. (2016). Mean absolute percentage error for regression models. *Neurocomputing*, *192*, 38–48.
- Demetrides, A., Smith, D. B., & Wang, X. (2018). General concepts of geochemical mapping at global, regional, and local scales for mineral exploration and environmental purposes. *Geochimica Brasiliensis*, *32*(2), 136–179.
- Dredge, L., & McMartin, I. (2011). Glacial stratigraphy of northern and central Manitoba. *Geological Survey of Canada, Bulletin*, *600*, 27. <https://doi.org/10.4095/288561>.
- Fix, E., & Hodges, J. L. (1951). Discriminatory analysis, non-parametric discrimination: Consistency properties. Technical Report 4, USAF School of Aviation Medicine, Randolph Field.
- Freund, Y., & Schapire, R. E. (1997). A decision-theoretic generalization of on-line learning and an application to boosting. *Journal of computer and system sciences*, *55*(1), 119–139.
- Friske, P. W. B. (1991). The application of lake sediment geochemistry in mineral exploration. In J. M. Franklin, J. M. Duke, W. W. Shilts, W. B. Coker, P. W. B. Friske, Y. T. Maurice, S. B. Ballantyne, C. E. Dunn, G. E. M. Hall & R. G. Garrett (Eds.), *Exploration Geochemistry Workshop* (pp. 157–180). Geological Survey of Canada, Open File 2390. <https://doi.org/10.4095/132392>.
- Friske, P. W. B., & Hornbrook, E. H. W. (1991). Canada's national geochemical reconnaissance programme. *Transactions of the Institution of Mining and Metallurgy, Section B. Applied Earth Science*, *100*, B47–B56.
- Garrett, R. G. (1983). Sampling methodology. In R. J. Howarth (Ed.), *Statistics and data analysis in geochemical prospecting* (Vol. 2, pp. 83–100). Elsevier Science. <https://doi.org/10.1016/B978-0-444-42038-1.50010-1>.
- Ghorbani, Y., Zhang, S. E., Nwaila, G. T., & Bourdeau, J. E. (2022). Framework components for data-centric dry laboratories in the minerals industry: A path to science-and-technology-led innovation. *The Extractive Industries and Society*, *10*, 101089.
- Ghorbani, Y., Zhang, S. E., Nwaila, G. T., Bourdeau, J. E., Safari, M., Hoseinie, S. H., Nwaila, P., & Ruuska, J. (2023). Dry laboratories—Mapping the required instrumentation and infrastructure for online monitoring, analysis, and characterization in the mineral industry. *Minerals Engineering*, *191*, 107971.
- Government of Canada. (2022). *The Canadian critical minerals strategy: from exploration to recycling* (p. 58). Government of Canada Natural Resources Canada.
- Govett, G. J. S. (1983). Geochemistry in the exploration sequence. In G. J. S. Govett (Ed.), *Handbook of Exploration Geochemistry* (Vol. 3, pp. 7–15). Elsevier. <https://doi.org/10.1016/B978-0-444-42021-3.50006-9>.
- Grunsky, E. C. (2010). The interpretation of geochemical survey data. *Geochemistry: Exploration, Environment, Analysis*, *10*(1), 27–74.
- Grunsky, E. C., & de Caritat, P. (2020). State-of-the-art analysis of geochemical data for mineral exploration. *Geochemistry: Exploration, Environment, Analysis*, *20*(2), 217–232.
- Grunsky, E. C., Mueller, U. A., & Corrigan, D. (2014). A study of the lake sediment geochemistry of the Melville Peninsula using multivariate methods: application for predictive geological mapping. *Journal of Geochemical Exploration*, *141*, 15–41.
- Hanmer, S., Sandeman, H. A., Davis, W. J., Aspler, L. B., Rainbird, R. H., Ryan, J. J., Relf, C., & Peterson, T. D. (2004). Geology and Neoproterozoic tectonic setting of the central Hearne supracrustal belt, western Churchill Province, Nunavut. *Canada. Precambrian Research*, *134*(1–2), 63–83.
- Harris, J. R., Grunsky, E., Behnia, P., & Corrigan, D. (2015). Data-and knowledge-driven mineral prospectivity maps for Canada's North. *Ore Geology Reviews*, *71*, 788–803.
- He, Y., Zhou, Y., Wen, T., Zhang, S., Huang, F., Zou, X., Ma, X., & Zhu, Y. (2022). A review of machine learning in geochemistry and cosmochemistry: Method improvements and applications. *Applied Geochemistry*, *140*, 105273.
- Ho, T. K. (1995). Random decision forests. In: *Proceedings of the 3rd International Conference on Document Analysis and Recognition* (pp. 278–282). Montréal. <https://doi.org/10.1109/ICDAR.1995.598994>.
- Hoffmann, J., Zortea, M., de Carvalho, B., & Zadrozny, B. (2021). Geostatistical Learning: Challenges and Opportunities. *Frontiers in Applied Mathematics and Statistics*, *7*, 689393.
- Hoffman, P. F. (1988). United plates of America, birth of a craton: Early Proterozoic assembly and growth of Laurentia. *Annual Reviews of Earth and Planetary Sciences*, *16*, 543–603.
- Jowitt, S. M., Mudd, G. M., Werner, T. T., Weng, Z., Barkoff, D. W., & McCaffrey, D. (2018). The critical metals, an overview and opportunities and concerns for the future. In A. M. Arribas & J. L. Mauk (Eds.), *Metals, minerals and society* (Vol. 21, pp. 25–38). Society of Economic Geologists. <https://doi.org/10.5382/SP.21.02>.
- Karpatne, A., Ebert-Uphoff, I., Ravela, S., Babaie, H. A., & Kumar, V. (2018). Machine learning for the geosciences: Challenges and opportunities. *IEEE Transactions on Knowledge and Data Engineering*, *31*(8), 1544–1554.
- Karrow, P. F., & Geddes, R. S. (1987). Drift carbonate on the Canadian Shield. *Canadian Journal of Earth Sciences*, *24*, 365–369.
- Kirkwood, C., Cave, M., Beamish, D., Grebby, S., & Ferreira, A. (2016). A machine learning approach to geochemical mapping. *Journal of Geochemical Exploration*, *167*, 49–61.
- Kotsiantis, S. B. (2014). Integrating global and local application of naive bayes classifier. *The International Arab Journal of Information Technology*, *11*(3), 300–307.
- Lary, D. J., Alavi, A. H., Gandomi, A. H., & Walker, A. L. (2016). Machine learning in geosciences and remote sensing. *Geoscience Frontiers*, *7*(1), 3–10.
- Lawley, C. J., Mitchell, M. G., Stralberg, D., Schuster, R., McIntire, E., & Bennett, J. R. (2022). Mapping Canada's green economic pathways for battery minerals: Balancing prospectivity modelling with conservation and biodiversity values. *Earth Science, Systems and Society*, *2*, 10064.
- Lawley, C. J. M., Tschirhart, V., Smith, J. W., Pehrsson, S. J., Schetselaar, E. M., Schaeffer, A. J., Houlié, M. G., & Eglinton, B. M. (2021). Prospectivity modeling of Canadian magmatic Ni (\pm Cu \pm Co \pm PGE) sulphide mineral systems. *Ore Geology Reviews*, *132*, 103985.
- Li, A. H., & Bradic, J. (2015). Boosting in the presence of outliers: adaptive classification with non-convex loss functions. *Journal of the American Statistical Association: Theory and Methods*. <https://doi.org/10.48550/1510.01064>.

- Linardatos, P., Papastefanopoulos, V., & Kotsiantis, S. (2021). Explainable AI: A review of ML interpretability methods. *Entropy*, 23(1), 1–18.
- MacHattie, T. G. (2001). Petrogenesis of the Wathaman batholith and La Ronge domain plutons in the Reindeer Lake area, Trans-Hudson Orogen, Saskatchewan. M.Sc. thesis, Memorial University, St. John's, Canada.
- Manitoba Mineral Resources, (2013). *Bedrock geology and mineral occurrences of Manitoba*. Map Gallery—Geoscientific Maps. Manitoba Mineral Resources. Retrieved January 20, 2023, from https://rdmaps.gov.mb.ca/Html5Viewer/index.html?viewer=MapGallery_Geology.MapGallery.
- Manitoba Natural Resources and Northern Development, (2021c). *Lithium*. Manitoba Natural Resources and Northern Development. Retrieved January 28, 2023, from https://www.gov.mb.ca/iem/geo/commodity/files/comm_lithium.pdf.
- Manitoba Natural Resources and Northern Development, (2021a). *Nickel*. Manitoba Natural Resources and Northern Development. Retrieved January 27, 2023, from https://www.gov.mb.ca/iem/geo/commodity/files/comm_nickel.pdf.
- Martins, T., Rayner, N., Corrigan, D., & Kremer, P. (2021). Regional geology and tectonic framework of the Southern Indian domain, Trans-Hudson Orogen, Manitoba. *Canadian Journal of Earth Sciences*, 59, 371–388.
- McCaffrey, K. J. W., Jones, R. R., Holdsworth, R. E., Wilson, R. W., Clegg, P., Imber, J., Holliman, N., & Trinks, I. (2005). Unlocking the spatial dimension: digital technologies and the future of geoscience fieldwork. *Journal of the Geological Society*, 162(6), 927–938.
- McCurdy, M. W., Amor, S. D., & Finch, C. (2016). Regional lake sediment and water geochemical data, Western and Central Labrador (NTS 13-L, 13-M, 14-D, 23-I, and 23-J). *Geological Survey of Canada, Open File*, 8026, 14. <https://doi.org/10.4095/298834>.
- McCurdy, M. W., & Garrett, R. G. (2016). Geochemical data quality control for soil, till and lake and stream sediment samples. *Geological Survey of Canada, Open File*, 7944, 40. <https://doi.org/10.4095/297562>.
- Mena Silva, C., Sørensen, B. E., Aasly, K., & Ellefmo, S. L. (2018). Geometallurgical approach to the element-to-mineral conversion for the Nabbaren nepheline syenite deposit. *Minerals*, 8(8), 325.
- Michaux, S. P. (2021a). The mining of minerals and the limits to growth. *Geological Survey of Finland, GTK Open File Work Report*, 16(2021), 73.
- Michaux, S. P. (2021b). Assessment of the extra capacity required of alternative energy electrical power systems to completely replace fossil fuels. *Geological Survey of Finland, GTK Open File Work Report*, 42(2021), 1000.
- Mudd, G. M. (2020). The resources cycle: key sustainability issues for the mining of metals and minerals. In: D. Alterton, S.A. Elias (Eds.), *Encyclopedia of Geology* (2nd ed., pp. 607–620). <https://doi.org/10.1016/B978-0-08-102908-4.00035-7>.
- Powell, K. (2019). *Far resources completes option on their Green Bay Property, Manitoba*. Canadian Mining and Energy. Retrieved January 26, 2023, from https://www.miningandenergy.ca/mininginsider/article/far_resources_completes_option_on_their_green_bay_property_manitoba/.
- Prior, T., Giurco, D., Mudd, G., Mason, L., & Behrisch, J. (2012). Resource depletion, peak minerals and the implications for sustainable resource management. *Global Environmental Change*, 22(3), 577–587.
- Manitoba Natural Resources and Northern Development, (2021b). *Rare Metals*. Manitoba Natural Resources and Northern Development. Retrieved January 28, 2023, from https://www.gov.mb.ca/iem/geo/commodity/files/comm_raremetals.pdf.
- Rodriguez-Galiano, V., Sanchez-Castillo, M., Chica-Olmo, M., & Chica-Rivas, M. J. O. G. R. (2015). Machine learning predictive models for mineral prospectivity: An evaluation of neural networks, random forest, regression trees and support vector machines. *Ore Geology Reviews*, 71, 804–818.
- Sabins, F. F. (1999). Remote sensing for mineral exploration. *Ore geology reviews*, 14(3–4), 157–183.
- Sagi, O., & Rokach, L. (2018). Ensemble learning: a survey. *Wiley Interdisciplinary Reviews Data Mining and Knowledge Discovery*, 8(4), e1249.
- Saskatchewan Geological Survey (2018). Resource Map of Saskatchewan. *Saskatchewan Ministry of the Economy, Saskatchewan Geological Survey, Miscellaneous report 2018-1*.
- Shirmard, H., Farahbakhsh, E., Müller, R. D., & Chandra, R. (2022). A review of machine learning in processing remote sensing data for mineral exploration. *Remote Sensing of Environment*, 268, 112750.
- Snow Lake Lithium (2022). Snow Lake Lithium receives \$158,000 grant from Manitoba chamber of commerce to support continued resource expansion. Retrieved January 26, 2023, from <https://ir.snowlakelithium.com/news-events/press-releases/detail/50/snow-lake-lithium-receives-158000-grant-from-manitoba>.
- Srivastava, P. K., Malhi, R. K. M., Pandey, P. C., Anand, A., Singh, P., Pandey, M. K., & Gupta, A. (2020). Revisiting hyperspectral remote sensing: Origin, processing, applications and way forward. *Hyperspectral remote sensing* (pp. 3–21). Elsevier.
- Stauffer, M. R. (1984). Manikewan and early Proterozoic Ocean in central Canada, its igneous history and orogenic closure. *Precambrian Research*, 25, 257–281.
- Sun, T., Chen, F., Zhong, L., Liu, W., & Wang, Y. (2019). GIS-based mineral prospectivity mapping using machine learning methods: A case study from Tongling ore district, eastern China. *Ore Geology Reviews*, 109, 26–49.
- Tarantola, A. (1984). Inversion of seismic reflection data in the acoustic approximation. *Geophysics*, 49(8), 1259–1266.
- Tschirhart, V., Pehrsson, S., Wodicka, N., Percival, J. A., Jefferson, C. W., Peterson, T., & Berman, R. G. (2022). Geophysical contributions to a synthesis of western Churchill geology and metallogeny. *Geological Survey of Canada, Preprint*. <https://doi.org/10.4095/330639>.
- Witten, I. H., & Frank, E. (2005). *Data mining: Practical machine learning tools and techniques* (2nd ed.). Morgan Kaufmann Publisher, Burlington.
- Zhang, C., Zuo, R., Xiong, Y., Zhao, X., & Zhao, K. (2022a). A geologically-constrained deep learning algorithm for recognizing geochemical anomalies. *Computers & Geosciences*, 162, 105100.
- Zhang, S. E., Bourdeau, J. E., Nwaila, G. T., & Corrigan, D. (2021a). Towards a fully data-driven prospectivity mapping methodology: A case study of the southeastern Churchill Province, Québec and Labrador. *Artificial Intelligence in Geosciences*, 2, 128–147.
- Zhang, S. E., Bourdeau, J. E., Nwaila, G. T., & Ghorbani, Y. (2022b). Advanced geochemical exploration knowledge using machine learning: Prediction of unknown elemental concentration and operational prioritization of re-analysis campaigns. *Artificial Intelligence in Geosciences*, 3, 86–100.
- Zhang, S., Carranza, E. J. M., Xiao, K., Chen, Z., Li, N., Wei, H., Xiang, J., Sun, L., & Xu, Y. (2021c). Geochemically constrained prospectivity mapping aided by unsupervised cluster analysis. *Natural Resources Research*, 30, 1955–1975.
- Zhang, S. E., Nwaila, G. T., Bourdeau, J. E., & Ashwal, L. D. (2021b). Machine learning-based prediction of trace element concentrations using data from the Karoo large igneous province and its application in prospectivity mapping. *Artificial Intelligence in Geosciences*, 2, 60–75.
- Zhang, S. E., Nwaila, G. T., Bourdeau, J. E., Ghorbani, Y., & Carranza, E. J. M. (2023). Deriving big geochemical data from high-resolution remote sensing data via machine

- learning: Application to a tailing storage facility in the Witwatersrand goldfields. *Artificial Intelligence in Geosciences*, 4, 9–21.
- Zuo, R. (2011). Identifying geochemical anomalies associated with Cu and Pb–Zn skarn mineralization using principal component analysis and spectrum–area fractal modeling in the Gangdese Belt, Tibet (China). *Journal of Geochemical Exploration*, 111(1–2), 13–22.
- Zuo, R., Luo, Z., Xiong, Y., & Yin, B. (2022). A geologically constrained variational autoencoder for mineral prospectivity mapping. *Natural Resources Research*, 31(3), 1121–1133.
- Zuo, R., Xiong, Y., Wang, J., & Carranza, E. J. M. (2019). Deep learning and its application in geochemical mapping. *Earth Science Reviews*, 192, 1–14.
- Zuo, R., & Xu, Y. (2023). Graph deep learning model for mapping mineral prospectivity. *Mathematical Geosciences*, 55(1), 1–21.
- Zuo, R., Zhang, Z., Zhang, D., Carranza, E. J. M., & Wang, H. (2015). Evaluation of uncertainty in mineral prospectivity mapping due to missing evidence: a case study with skarn-type Fe deposits in Southwestern Fujian Province, China. *Ore Geology Reviews*, 71, 502–515.